A Calibrated and Interpretable Framework for Multilingual Text Difficulty Prediction

Voula Giouli, George Tsoulouhas, Athina Sioupi, Stamatia Michalopoulou
Aristotle University of Thessaloniki
Thessaloniki, Greece
pgiouli@del.auth.gr, george.tsoulouhas@athenarc.gr, sioupi@del.auth.gr, smichalo@del.auth.gr

Abstract

We present a framework for automatic text difficulty prediction, centered on a linguistically enriched German dataset comprising texts that are aligned with the levels defined by the Common European Framework of Reference for Languages, with its Greek counterpart currently under development. The dataset bears annotations and is used for training and evaluating a system that integrates: (i) a lexicon of lexical profiling enriched via existing openly available lexical resources, (ii) large-scale frequency modeling using a 1.5M-word corpus, (iii) syntactic complexity pre-computation, (iv) percentile-based rule calibration, and (v) feature-based machine learning classifiers with feature selection. Our framework integrates BERT-based contextual modeling and augments it with SHAP-driven interpretability mechanisms and rigorous analysis of confidence calibration. The framework is designed to be in principle language-agnostic, aiming to facilitate multilingual portability.

Keywords: readability prediction, text difficulty, Computer-Assisted Language Learning

1. Introduction

Automatic text difficulty prediction is central to computer-assisted language learning (CALL), adaptive educational systems, curriculum development, and automated content recommendation. In European contexts, text difficulty is typically assessed through the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) which defines six proficiency levels (A1-C2). Although recent research increasingly adopts transformer-based architectures, educational applications require interpretability and pedagogical justification. This is particularly true in language learning contexts, where model outputs must be transparent, explainable and aligned with established proficiency standards.

This paper presents work aimed at developing a workbench for CEFR-based text difficulty prediction. The proposed platform comprises three main components: (i) a tool for CEFR-aligned dataset preparation incorporating a pipeline for documenting, processing, and enriching textual data, (ii) CEFR-aligned datasets, and (iii) three alternative modeling approaches, namely a rule-based baseline, a feature-based Machine Learning (ML) classifier and a fine-tuned BERT model. Our approach integrates linguistically informed feature engineering with data-driven modeling techniques, thereby balancing transparency and predictive performance.

The proposed workbench has been designed within the EmoBot project (Kallipolitis et al., 2026) as a language-agnostic infrastructure that can be extended to any language. In its current implementation, it has been applied to the creation of a German CEFR dataset, while its Greek counterpart is currently under development.

This work makes three main contributions: (i) an extensible infrastructure for building interoperable CEFR-aligned corpora in multiple languages; (ii) a CEFR-annotated German dataset with a Greek one currently underway to evaluate portability and extensibility; and (iii) a comprehensive text difficulty prediction framework that integrates a calibrated rule-based approach, a ML classifier, and BERT modelling for CEFR-based text difficulty assessment, currently evaluated on German.

The paper is structured as follows: Section 2 reviews related work on text difficulty assessment. The infrastructure, including the CEFR-annotation tool, the CEFR-aligned dataset and lexicon, is described in Section 3, while Section 4 details the predictive modeling approaches. Sections 5 and 6 present and discuss the empirical findings. We conclude and outline future research in Section 7.

2. Related work

Readability assessment - defined as the degree of "legibility, interest, or ease of reading" associated with a text (Dale and Chall, 1949) - has long been a central topic in applied linguistics, education, and computational language processing. In language pedagogy, automated readability assessment is valuable not only for native speakers but is particularly important for second/foreign language (L2/FL) learners, as it supports the systematic alignment of texts with learners' proficiency levels and instructional objectives. In European educational contexts, CEFR provides a standardized scale for describing language proficiency and guiding curriculum

Proceedings of the 2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2026), pages 63–73

design.

Early approaches to text difficulty assessment relied on surface-level readability formulas designed to approximate cognitive processing difficulty using shallow linguistic indicators. These formulas typically combine sentence and word length or syllable count to estimate educational grade level, as for example, the Flesch Reading Ease score (Flesch, 1948) and its derivative, the Flesch–Kincaid Grade Level (Kincaid et al., 1975). However, they have been proved to be useful for larger pieces of language (texts) rather than shorter texts or dialogues usually used in current educational CALL applications (Rooein et al., 2024).

Additional formula-based indices such as the Coleman–Liau Index (Coleman and Liau, 1975) and the Automated Readability Index (ARI) (Smith and Senter, 1967) rely on character-level metrics rather than syllable counts, making them easier to compute programmatically.

Language-specific formulas e.g., for German, namely the Wiener Sachtextformel (Bamberger and Vanecek, 1984) are tailored to specific linguistic properties. Readability formulas, however, exhibit several limitations in that they rely on shallow surface features rather than explicitly modeling syntactic or discourse structure. From an educational point of view, these formulas are not directly aligned with CEFR descriptors while they generalize poorly across typologically distinct languages without adaptation.

Over time, the field has expanded to incorporate lexical frequency, syntactic complexity, discourse structure, and, more recently, machine learning and neural modeling approaches.

2.1. Machine Learning approaches

Supervised feature-based models were among the first to be used in this direction. Vajjala and Meurers (2012) advanced readability classification by incorporating insights from L2 acquisition research, combining measures of lexical richness with syntactic complexity features (including parse-tree-based ones), and demonstrating significant improvements over traditional formulas through experiments with Support Vector Machines (SVMs). For German specifically, Hancke et al. (2012) developed a readability classifier using lexical, syntactic, and morphological features on a graded corpus of school and web texts, establishing an early benchmark for German text difficulty assessment.

SVM classifiers are very common in CEFR-based classification. Xia et al. (2016) attained 80.3% accuracy on the WeeBit corpus utilizing an SVM classifier with lexical and syntactic features supplemented by discourse indicators. Meanwhile, Pilán et al. (2016) illustrated that weakly lexicalized features improve generalization across unseen data for CEFR-level assessment in Swedish L2 reading materials. In addition, Random Forest classifiers have shown strong results in text classification tasks, and Imperial et al. (2025) reported good results for document-level CEFR classification.

The emergence of neural network methodologies has established novel paradigms in readability evaluation. Azpiazu and Pera (2019) proposed Multiattentive Recurrent Neural Networks for multilingual readability prediction, achieving 84.7% accuracy on the VikiWiki dataset. More recently, transformer-based architectures have shown promising results: Deutsch et al. (2020) showed that BERT embeddings enhance readability prediction beyond traditional feature-based methods alone, while Imperial (2021) demonstrated that hybrid models combining BERT sentence embeddings with handcrafted linguistic features typically outperform either approach in isolation. Fine-tuned BERT models have also proven effective for direct CEFR classification, as evidenced by Santos et al. (2021). A consistent theme across studies is that feature-based models provide superior interpretability and robustness with limited training data, whereas transformer-based models deliver higher discriminative performance when ample data are available (Martinc et al., 2021).

2.2. LLMs for Text Difficulty Assessment

The emergence of LLMs has opened new pathways for assessing text difficulty. Yancey et al. (2023) evaluated GPT-4 for rating short L2 essays on the CEFR scale; few-shot calibration was found to approach the accuracy of state-of-the-art automated writing evaluation systems, though agreement with human ratings varied by learners' native languages. Trott and Rivière (2024) showed that GPT-4 Turbo's zero-shot readability estimates are very similar to human judgments (r = 0.76) and work better than traditional formulas on the CLEAR corpus.

While LLM-based approaches show clear performance gains—Rooein et al. (2024) demonstrated that prompt-based LLM metrics improve difficulty classification over traditional measures such as Flesch-Kincaid, particularly for shorter texts—measurement stability remains a concern. Uchida (2024) reported inconsistencies in ChatGPT-generated CEFR ratings in repeated evaluations, suggesting that single-pass LLM assessments may lack the reproducibility required for high-stakes educational applications. Recent research on the Ace-CEFR dataset (Kogan et al., 2025) provides a thorough assessment, contrasting linear models, BERT-based classifiers, and LLM-based methodologies. The results demonstrate that hybrid techniques, which integrate BERT embeddings with LLM scoring, can surpass the performance of individual human expert raters. Nevertheless, LLM

inference is substantially slower than feature-based or BERT-based classification, rendering it more suitable for offline labeling; this efficiency trade-off, coupled with the limited transparency of LLM-internal reasoning, highlights the ongoing value of interpretable feature-based approaches in educational contexts where pedagogical justification is essential.

3. Text Difficulty Assessment Workbench

In this section, we present the infrastructure developed, namely the tool for (semi-) automatically creating a CEFR-aligned corpus, the CEFR-aligned dataset that has been developed as a proof-of-concept and the tools for text difficulty prediction.

3.1. CEFR-annotation tool

The annotation tool is a web-based platform designed to support the full lifecycle of CEFR-aligned corpus construction, from text ingestion and collaborative annotation to quality assurance and multi-format export. The tool is built around a modular architecture comprising several interconnected subsystems described below.

Text management and metadata annotation. The tool provides a structured interface for creating, editing, and managing textual entries. Each text record stores the raw content alongside rich metadata including the assigned CEFR level (A1–C2), genre classification, thematic domain, register, and source attribution. Selected texts follow a status-based workflow progressing through draft, pending review, approved, and rejected stages, ensuring that only quality-controlled entries enter the final corpus. Upon ingestion, texts are automatically analyzed to extract a comprehensive set of linguistic features, including readability indices, lexical diversity measures, and CEFR-aligned vocabulary profiles, which are stored as structured metadata for downstream use in both rule-based and machine learning pipelines.

Text segmentation. Approved texts exceeding a minimum length can be split into smaller segments through a manual segmentation interface. Segments are stored as child records linked to the parent text, preserving the parent's core metadata while maintaining independent CEFR annotations and segment ordering. This feature supports the creation of discourse-level sub-corpora from longer documents.

Activity creation and management. The tool supports the creation of structured CEFR-leveled language learning activities linked to annotated texts.

Role-based access control and annotation workflow. The tool implements a role-based permission system with three user roles: administrator, reviewer, and annotator. Administrators have full system access including user management and model training (see Section 4.2.2). Reviewers manage the annotation workflow by assigning texts to annotators, reviewing submitted assessments, and setting final CEFR levels upon approval. Annotators provide independent CEFR assessments on assigned texts, with the system hiding existing level assignments to prevent bias. This multi-stage workflow — creation, assignment, independent assessment, and expert review — is designed to produce reliable annotations suitable for supervised learning.

Inter-annotator agreement. To assess annotation reliability, the tool computes inter-annotator agreement statistics. It employs Cohen's Kappa for pairwise comparisons and Fleiss' Kappa when three or more annotators assess the same text. Weighted Kappa with linear weights is used to account for the ordinal nature of the CEFR scale. Agreement scores are interpreted according to the Landis and Koch scale and are displayed alongside per-text annotation summaries, enabling reviewers to identify texts with divergent assessments and prioritize them for adjudication.

Statistics and analytics dashboard. A dedicated statistics dashboard provides aggregated views of the corpus, including CEFR-level distributions for texts, words, and sentences; content distributions by genre, topic, and register; quality metrics per proficiency level (average word count, readability scores, lexical diversity); inter-annotator agreement summaries; and a confusion matrix comparing predicted and final CEFR levels. Per-annotator performance metrics are also available. All statistics can be exported to a formatted Excel workbook.

Multi-format corpus export. Beyond ML-oriented dataset generation, the tool supports corpus export in multiple standard formats: JSON, CSV, TSV, XML, RDF/OWL (using Dublin Core metadata terms and a custom CEFR namespace), formatted Excel workbooks, and the Universal-CEFR (Imperial et al., 2025) JSON schema for interoperability with shared CEFR datasets. Exports can be filtered by CEFR level, annotation status, genre, register, topic, and dataset split. Both full exports (including prediction metadata) and minimal variants are available.

REST API and programmatic access. The tool exposes a REST API for programmatic interaction. Key endpoints include CEFR prediction (combining rule-based and machine learning outputs with full feature breakdowns), text and activity export with filtering, and corpus statistics retrieval.

3.2. CEFR-aligned dataset

For the purposes of the present study, a structured and pedagogically oriented corpus of authentic written texts was systematically compiled from a broad spectrum of digital sources, including newspapers, general-interest magazines, and electronic media outlets operating within the German-speaking context (e.g., television networks providing written news reports). The inclusion criteria were guided by principles of authenticity, communicative relevance, and representativeness of contemporary language use. Only openly accessible sources were considered.

In order to ensure alignment with standardized language assessment practices and established proficiency benchmarks, the corpus was supplemented with sample examination materials issued by officially recognized certification bodies. These included open access materials from the State Certificate of Language Proficiency in Greek (KPG), as well as publicly available sample papers provided by the Goethe Institut and the Österreichisches Sprachdiplom Deutsch (ÖSD) for proficiency levels A–C. Furthermore, pedagogically graded resources *from the platform of Deutsche Welle* were systematically integrated, given their explicit alignment with proficiency descriptors and their didactic scaffolding.

The classification and annotation of the texts with respect to thematic domains and genre typology were conducted in accordance with the proficiency levels and thematic specifications defined for each level (A1–C2) by the CEFR. The categorization process followed a level-sensitive and descriptor-informed approach, ensuring coherence between linguistic complexity, communicative function, textual genre, and thematic scope. The corpus encompasses a diverse range of genres, including advertisements, journalistic articles, blog posts, and dialogic interactions, not only from examination material but also from online newspapers and magazines. The thematic spectrum is correspondingly broad, covering domains such as biographical narratives, education, culture, and entertainment, thereby facilitating exposure to varied discourse types, registers, and communicative contexts. Annotation was performed by two trained linguists followed by a review/adjudication process. Pairwise inter-annotator agreement yielded a linearly weighted Cohen's κ of 0.75, indicating substantial agreement on the Landis and Koch scale (Landis and Koch, 1977).

Currently, the dataset comprises a set of CEFR-aligned texts along with CEFR-leveled activities. In specific, the dataset amounts to c.920 texts, 854K tokens and 59K sentences. Each text in the corpus is stored together with its raw textual content, assigned CEFR level, a comprehensive set of linguistic feature metadata, pre-computed syntactic complexity measures, and a CEFR-aligned vocabulary profile. In terms of size and depth of linguistic annotation, the dataset represents one of the more extensive manually curated CEFR-aligned German resources currently available for research in automated text difficulty assessment.

The CEFR-labeled subset covers all six proficiency levels (A1–C2). Although the number of texts is distributed across levels, advanced stages (C1 and C2) account for a substantial proportion of the total word count, reflecting the greater length, lexical density, and structural complexity characteristic of higher-level materials. This distribution supports the modeling of proficiency progression from beginner to intermediate levels, provides a rich representation of advanced syntactic phenomena, and enables systematic analysis of confusion patterns between adjacent CEFR levels. Table 1 gives an overview of the CEFR-aligned corpus.

CEFR	TC	AvgWC	AvgR	AvgLDiv	AvgSL
A1	148	77.81	73.60	0.7804	7.71
A2	170	158.36	64.36	0.7463	9.79
B1	164	162.40	58.13	0.7617	11.89
B2	155	245.10	52.41	0.7359	14.44
C1	187	719.82	49.39	0.6169	17.25
C2	97	1061.45	43.92	0.5864	18.05

Table 1: Descriptive statistics per CEFR level for the German corpus: Text Count (TC), Average Word Count (AvgWC), Average Readability (AvgR; Flesch Reading Ease, Amstad German adaptation), Average Lexical Diversity (AvgLDiv; type–token ratio), and Average Sentence Length (AvgSL, in tokens).

3.3. Word complexity lexicon

The lexical backbone of the system consists of a structured resource centered on a comprehensive German word list that depicts lexical complexity in terms of CEFR levels. As a starting point, we used officially published CEFR-aligned word lists as seeds, ensuring reliable level assignments for core vocabulary. These seed lists were imported via the DWDS API[1] This resource enables the computation of CEFR-specific vocabulary distributions within texts (A1–C2 percentages) as well as derived indicators such as basic-to-advanced vocabulary ratios.

Since official lists provide limited coverage at higher proficiency levels, we implemented a frequency-based expansion strategy using approximately 1.5M German lemmas derived from the OpenSubtitles corpus (Lison and Tiedemann,

[1] https://www.dwds.de/d/api

2016)[2] ranked by frequency[3]. OpenSubtitles was selected for its large scale and broad genre coverage; however, its conversational register may underrepresent academic and formal language typical of higher CEFR levels. The resulting frequency-based CEFR assignments should therefore be understood as corpus-derived approximations of proficiency level rather than pedagogically validated CEFR annotations in the strict sense. The raw frequency ranks were normalized to a 1–1000 scale to ensure uniform comparability and computational stability. We then partitioned the ranked vocabulary into frequency bands aligned with CEFR levels, assigning the 1–2,000 most frequent items to A1, 2,000–5,000 to A2, 5,000–10,000 to B1, 10,000–20,000 to B2, 20,000–35,000 to C1, and items beyond 35,000 to C2. This stratification reflects the well-established correlation between lexical frequency, acquisition sequence, and proficiency development, thereby providing a principled mechanism for approximating CEFR levels for out-of-list lexical items. While this mapping is heuristic rather than prescriptive, it offers a scalable and corpus-derived baseline that can be recalibrated as additional CEFR-annotated data become available.

To incorporate morphosyntactic complexity beyond frequency, lexical entries are enriched with information extracted from Wiktionary (Wiktionary contributors, 2024), including irregular verb status, separable verb properties, compound formation, and international word marking. These annotations capture structural properties associated with acquisition difficulty in German and provide linguistically interpretable signals for modeling. We note that the effect of internationalisms on perceived difficulty is L1-dependent: cognates may facilitate comprehension for learners with Romance L1 backgrounds while offering less support to learners from typologically distant language families. In the current implementation, international word status is treated as a uniform feature; future work could condition its contribution on learner L1. The resulting vocabulary resource directly feeds into the machine learning pipeline.

4. Text difficulty assessment

4.1. Syntactic complexity modeling

All texts undergo automatic dependency parsing using spaCy (Honnibal et al., 2020) models, allowing the extraction of measures of syntactic complexity on a scale. Pre-computed features include subordinate, relative, and infinitive clause ratios, passive voice frequency, modal verb density, nominalization ratio, clause embedding depth, average dependency distance, and noun–verb ratio. These indicators capture structural phenomena associated with proficiency progression, particularly the increasing use of clause embedding and nominal structures at higher CEFR levels. Syntactic features are computed once and stored in the database, ensuring reproducibility and efficient retraining without repeated parsing. This pre-analysis design reduces computational overhead during model experimentation and facilitates integration with both rule-based and machine learning approaches. Together with lexical and readability measures, the syntactic layer provides complementary structural signals that improve level discrimination, as shown in the confusion analysis in Section 6.

4.2. Baseline and Machine Learning

4.2.1. Rule-Based Baseline

As a transparent and pedagogically interpretable baseline, we developed a rule-based scoring system that combines three independent prediction components: (a) CEFR-aligned vocabulary distribution analysis, (b) readability metric voting, and (c) linguistic complexity scoring.

The **vocabulary component** computes the cumulative CEFR level coverage for each text using the lexical resource described in Section 3.3. After lemmatization with spaCy and filtering of proper nouns, each token is assigned to its CEFR level. A text is assigned to a given level if the cumulative percentage of words at or below that level exceeds a threshold (e.g. $\geq$85% for A1, $\geq$80% for A2, decreasing to $\geq$65% for C1). Texts with more than 30% unknown vocabulary default to C2; this threshold serves as a fallback for out-of-vocabulary texts. In practice, unknown-word rates are much lower across all levels, with C2 texts averaging 8.1% and C1 texts 6.4%. Confidence is derived from the margin above the threshold, adjusted downward when the unknown-word percentage is high.

The **readability component** employs weighted voting across seven readability metrics: Wiener Sachtextformel (weight 1.0), Flesch Reading Ease in the Toni Amstad German adaptation (0.9), Flesch–Kincaid Grade Level (0.7), Gunning Fog Index (0.6), SMOG Index (0.6), Automated Readability Index (0.65) and Coleman–Liau Index (0.65). Each metric independently maps its score to a CEFR level using predefined range tables. The final readability prediction is determined by weighted majority voting, with confidence modulated by a spread factor that penalizes disagreement among metrics.

The **linguistic component** computes a composite score (0–100) that aggregates four feature cate-

[2] http://www.opensubtitles.org/

[3] https://github.com/hermitdave/FrequencyWords

gories: syntactic complexity (subordination depth, mean dependency distance, clauses per sentence), grammatical features (passive voice, subjunctive mood, genitive case, complex tenses), lexical sophistication (type–token ratio, average frequency rank, compound word ratio) and discourse markers (proportion of advanced and sophisticated connectives). The composite score is mapped to CEFR levels using calibrated boundaries.

The three components are combined via weighted index averaging. Default weights (vocabulary: 0.30, readability: 0.40, linguistic: 0.30) are dynamically adjusted based on vocabulary coverage quality and linguistic signal strength. Ensemble confidence is boosted when components agree and reduced when all three predict different levels, in which case the median prediction is selected.

Calibration is data-driven: using the approved subset of the corpus, the system extracts per-level feature distributions and computes percentile-based thresholds for syntactic features (e.g., mean dependency distance boundaries at 2.5, 2.75, 3.09, 3.38 for the A1/A2 through C1 transitions). Component weights are optimized based on individual prediction accuracy on the calibration set, yielding readability: 0.424, vocabulary: 0.31, linguistic: 0.266.

4.2.2. Feature-Based Machine Learning

Building on the linguistic features used in the rule-based system, we train supervised classifiers using an expanded feature set of 89 dimensions organized into seven categories: (1) basic text statistics, (2) readability metrics, (3) CEFR vocabulary profile, (4) derived vocabulary ratios, (5) morphosyntactic features, (6) syntactic complexity, and (7) TF-IDF features (see Appendix 11.1). All features are standardized using z-score normalization fitted on the training partition. We evaluate four classifiers: Random Forest (RF), Gradient Boosting (GB), Support Vector Machine with RBF kernel (SVM), and XGBoost. Class imbalance is addressed through balanced class weighting for RF and SVM. Model selection employs grid search with 3-fold stratified cross-validation, optimizing for macro F1. Data are split into training (70%), development (15%), and test (15%) partitions with stratified sampling.

Four feature selection strategies are available: importance-based filtering (retaining features with Random Forest importance ≥ 0.01), correlation-based filtering (removing features with Pearson $r > 0.85$), recursive feature elimination (RFE), and a combined approach applying correlation filtering followed by importance-based selection.

Model interpretability is supported through SHAP (SHapley Additive exPlanations) analysis (Lundberg and Lee, 2017). For tree-based models, exact SHAP values are computed via `TreeExplainer`; for SVM, approximate values are obtained via `KernelExplainer`. The system provides global feature importance rankings, per-class feature contributions, and local prediction explanations for individual texts.

Evaluation metrics include accuracy, macro and weighted F1 scores, per-class precision, recall, and F1, and adjacent accuracy (predictions within ± 1 CEFR level). Probability calibration is assessed via Expected Calibration Error (ECE), Maximum Calibration Error (MCE), and the multi-class Brier score.

4.3. BERT modelling

We fine-tune a German BERT model (`deepset/gbert-base`) (Chan et al., 2020) for direct CEFR classification. The model consists of 12 transformer layers with a hidden size of 768 dimensions. A classification head comprising a dropout layer ($p = 0.3$) followed by a linear projection ($768 \rightarrow 6$) is added on top of the `[CLS]` token representation.

Training uses the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 2×10^{-5}, weight decay of 0.01, and a linear warmup schedule (10% of total steps) followed by linear decay. Sequences are tokenized using WordPiece tokenization and truncated or padded to a maximum length of 512 tokens. Class imbalance is handled through inverse-frequency weighting in the cross-entropy loss, combined with label smoothing ($\epsilon = 0.1$) to mitigate overconfident predictions. Early stopping monitors macro F1 on the validation set with a patience of 3 epochs.

Mixed-precision training (FP16) is enabled on CUDA devices via PyTorch's Automatic Mixed Precision, with gradient clipping at a maximum norm of 1.0. The system automatically detects the available hardware, supporting NVIDIA GPUs (CUDA), Apple Silicon (MPS), and CPU fallback. Four training presets are provided to accommodate different computational budgets: a *default* configuration (batch size 8, 10 epochs, learning rate 2×10^{-5}), a *fast* preset (batch size 16, 5 epochs, max length 256), an *accurate* preset (batch size 4, 15 epochs, learning rate 1×10^{-5}, dropout 0.2, patience 5), and a *CPU-friendly* preset (batch size 4, max length 256, no mixed precision).

Data partitioning follows the same stratified 70/15/15 split as the feature-based models, ensuring comparable evaluation. At inference, the model outputs a softmax probability distribution over all six CEFR levels, enabling both point predictions and uncertainty-aware decision-making.

5. Results

We evaluated three modeling approaches on the German CEFR dataset described in Section 3.2: the rule-based baseline, the best-performing feature-based ML classifier, and the fine-tuned BERT model. The rule-based and ML models were assessed on the same held-out test partition ($n = 134$), and the BERT model on a comparable partition ($n = 136$), both using stratified sampling to preserve class distribution.

5.1. Rule-Based Classification

The rule-based system combines three signal sources through weighted voting: vocabulary profile analysis (weight 0.30), readability metrics from multiple formulas (weight 0.40), and linguistic feature analysis including syntactic complexity (weight 0.30). Weights are dynamically adjusted based on input quality—for instance, vocabulary weight increases when unknown-word percentage is low, indicating good dictionary coverage.

Table 2 presents the per-class performance. The system achieves an overall accuracy of 43.3% and a macro F1 of 0.385. Adjacent accuracy reaches 88.1%, indicating that most errors involve neighboring levels rather than large classification jumps.

Level	Precision	Recall	F1	Support
A1	0.833	0.526	0.645	19
A2	0.560	0.560	0.560	25
B1	0.439	0.720	0.545	25
B2	0.286	0.522	0.369	23
C1	0.286	0.143	0.190	28
C2	0.000	0.000	0.000	14
Macro	**0.401**	**0.412**	**0.385**	134

Table 2: Per-class performance of the rule-based classifier on the test set (accuracy: 43.3%, adjacent accuracy: 88.1%).

Performance is strongly skewed toward lower proficiency levels: A1 achieves the highest precision (0.833) and a reasonable F1 (0.645), while A2 and B1 reach F1 scores of 0.560 and 0.545 respectively. In contrast, the system struggles with higher levels—B2 drops to 0.369, C1 to 0.190, and C2 is never correctly predicted (F1 = 0.000). The confusion matrix reveals the primary failure mode: a systematic downward bias, where C1 texts are most frequently misclassified as B2 (21 out of 28), and all 14 C2 texts are assigned to levels B1–C1. This pattern reflects the limitations of readability formulas and vocabulary frequency analysis for distinguishing advanced proficiency levels, where textual complexity manifests through discourse-level features (argumentation structure, hedging, register) rather than surface-level statistics. These results establish a clear motivation for the machine learning approaches that follow: while the rule-based system provides a functional baseline for lower levels (A1–B1), it fails to discriminate among upper levels (B2–C2), precisely the range where learner placement decisions are most consequential.

5.2. Feature-Based Classification

We trained four classifiers using 3-fold stratified cross-validation with grid search over hyperparameter grids: Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), and XGBoost. Table 3 reports the cross-validation macro F1 and test set macro F1 for each model.

Classifier	CV F1	Test F1	Acc.	Adj.Acc.
RF	**0.602**	0.597	0.612	0.955
GB	0.567	**0.602**	0.612	0.963
SVM	0.576	0.588	0.597	0.978
XGBoost	0.601	0.600	0.612	0.963

Table 3: Classifier selection via 3-fold stratified cross-validation with grid search (macro F1). Random Forest is selected based on highest CV macro F1.

All four classifiers perform within a narrow range (test macro F1: 0.588–0.602), with Random Forest achieving the highest CV score (0.602) and Gradient Boosting the highest test F1 (0.602). Given this near-parity, we selected Random Forest as the final model based on its best cross-validation performance and its support for feature importance analysis. Its best hyperparameters are: `max_depth`=7, `n_estimators`=200, and `min_samples_split`=2.

Table 4 summarizes the performance of the feature-based classifier trained with 40 selected features (including syntactic and TF-IDF components). The model achieves an overall accuracy of 61.2% and a macro F1 of 0.597. Notably, adjacent accuracy reaches 95.5%, indicating that the vast majority of misclassifications involve adjacent CEFR levels — a pattern consistent with the inherent difficulty of fine-grained ordinal classification on a six-level scale.

Level	Precision	Recall	F1	Support
A1	0.722	0.684	0.703	19
A2	0.645	0.800	0.714	25
B1	0.579	0.440	0.500	25
B2	0.571	0.522	0.545	23
C1	0.606	0.714	0.656	28
C2	0.500	0.429	0.462	14
Macro	**0.604**	**0.598**	**0.597**	134

Table 4: Per-class performance of the feature-based ML classifier on the test set (accuracy: 61.2%, adjacent accuracy: 95.5%).

Performance is strongest at the lower end of the scale: A2 achieves the highest F1 (0.714), followed by A1 (0.703) and C1 (0.656), while C2 and B1 prove most challenging (F1 of 0.462 and 0.500, respectively). This pattern reflects the well-documented difficulty of discriminating between adjacent proficiency levels, where linguistic features overlap substantially.

Calibration analysis reveals an Expected Calibration Error (ECE) of 0.242 and a Brier score of 0.634. The model exhibits overconfidence: average confidence for incorrect predictions is 74.1% compared to 81.0% for correct ones, with 91.5% of errors made with confidence above 0.5. These findings underscore the importance of calibration-aware evaluation in educational applications where prediction confidence informs downstream decisions.

5.3. BERT-Based Classification

The fine-tuned German BERT model was trained using the *accurate* preset (batch size 4, 15 epochs, learning rate 1×10^{-5}, dropout 0.2). Training converged at epoch 14 based on validation macro F1 (0.708). Table 5 presents the test set performance.

Level	Precision	Recall	F1	Support
A1	0.800	0.909	0.851	22
A2	0.786	0.880	0.830	25
B1	0.636	0.583	0.609	24
B2	0.545	0.522	0.533	23
C1	0.600	0.556	0.577	27
C2	0.571	0.533	0.552	15
Macro	**0.656**	**0.664**	**0.659**	136

Table 5: Per-class performance of the fine-tuned BERT model on the test set (accuracy: 66.9%, adjacent accuracy: 95.6%). Rule-based and ML evaluated on n=134; BERT on n=136 due to different stratified split.

The BERT model achieves an accuracy of 66.9% and a macro F1 of 0.659, representing a relative improvement of 10.4% over the feature-based classifier. Adjacent accuracy reaches 95.6%. The improvement is most pronounced for A1–A2 levels, where BERT achieves F1 scores of 0.851 and 0.830 respectively. The B2 level proves most challenging (F1 = 0.533), followed by C2 and C1. B2 occupies a transitional zone between intermediate and advanced proficiency, sharing surface-level features (sentence length, vocabulary breadth) with both B1 and C1, which makes categorical discrimination particularly difficult.

6. Discussion

The progression from rule-based to BERT yields consistent gains: accuracy improves from 43.3% to 61.2% to 66.9%, while macro F1 rises from 0.385 to 0.597 to 0.659. The most dramatic improvement occurs in the upper CEFR levels where the rule-based system fails entirely—the ML classifier recovers C1 (F1 = 0.656) and C2 (F1 = 0.462), while BERT achieves 0.577 and 0.552 respectively. Table 6 summarizes the three approaches.

Metric	Rule-Based	ML (Feature)	BERT
Accuracy	0.433	0.612	0.669
Macro F1	0.385	0.597	0.659
Weighted F1	0.401	0.605	0.663
Adjacent Acc.	0.881	0.955	0.956
Interpretability	High	Medium	Low

Table 6: Comparison of the three modeling approaches on the held-out test split.

Confusion matrix analysis across all three models reveals a consistent pattern: misclassifications concentrate along the diagonal, predominantly between adjacent levels. The B1–B2 boundary is the most error-prone region, followed by C1–C2. A1 texts are the most reliably classified across all approaches, likely due to their distinctive short sentence length, limited vocabulary, and low syntactic complexity. The high adjacent accuracy across all models ($>$88%) suggests that the six-level CEFR scale introduces inherent ambiguity at level boundaries, and that the models capture the ordinal structure of the proficiency continuum even when exact-level prediction fails.

The rule-based baseline, while the least accurate as a standalone classifier, offers full transparency: every prediction can be traced to specific vocabulary thresholds, readability scores, and linguistic indicators. This interpretability makes it suitable for the semi-automatic annotation workflow described in Section 3.1, where human annotators use the rule-based prediction as a starting point. The feature-based ML classifiers provide a middle ground, achieving competitive performance while supporting SHAP-based post-hoc interpretability. SHAP analysis of the Random Forest model reveals that word count, the first TF-IDF principal component, and average sentence length are the three most influential features globally. Notably, the driving features shift across proficiency levels: at lower levels (A1–A2), vocabulary profile features (A1 vocabulary percentage, basic vocabulary ratio) dominate, while at advanced levels (C1–C2), text length and TF-IDF features become the primary discriminators, reflecting the greater lexical and structural diversity characteristic of higher-proficiency texts. BERT delivers the highest discriminative performance but operates as a black-box model, with prediction confidence as the primary transparency mechanism.

It is worth noting that none of the three ap-

proaches achieves particularly high exact-match accuracy, which may appear surprising at first glance. However, this outcome is entirely expected given the nature of CEFR level assignment. Many texts lie at the boundary between two adjacent levels, and human annotators must make a categorical decision where the underlying proficiency is continuous. This annotation process inherently introduces a degree of subjectivity: two expert annotators may reasonably disagree on whether a text is B1 or B2, for instance. The moderate exact-match accuracy across all three methods directly reflects this inter-annotator ambiguity in the training data. Crucially, the consistently high adjacent accuracy (88.1% for rule-based, 95.5% for ML, and 95.6% for BERT) demonstrates that when the models err, they almost always predict a neighboring level—mirroring the same boundary uncertainty that human annotators face.

Furthermore, the prediction confidence scores provide additional diagnostic value: when a model assigns a text to a given level with low confidence and high probability mass on an adjacent level, this signals that the text genuinely straddles two proficiency levels, offering practitioners actionable information beyond the categorical prediction alone.

7. Conclusion

We have presented a framework for CEFR-based text difficulty prediction that combines a transparent rule-based baseline, interpretable feature-based ML classifiers, and a fine-tuned BERT model. Evaluated on a 920-text German corpus, the three approaches yield progressively higher accuracy (43.3%, 61.2%, 66.9%) while adjacent accuracy exceeds 88% across all models, confirming that errors predominantly involve neighbouring CEFR levels. The B1–B2 boundary remains the most challenging region even for BERT, reflecting inherent ambiguity at mid-proficiency transitions. SHAP analysis reveals that the dominant predictive features shift from vocabulary profile at lower levels to text length and TF-IDF components at advanced levels, offering pedagogically interpretable insights into what distinguishes proficiency stages. In its current implementation, the framework handles German, with its Greek counterpart under development. The cross-linguistic extension will allow systematic investigation of feature transferability across typologically distinct languages.

Future work is planned in two directions: (a) finalising the Greek CEFR-aligned corpus and adapting the classifier for Modern Greek, and (b) leveraging LLMs for controlled text simplification, using the calibrated difficulty predictions and feature-level insights from the current framework as constraints to guide generation toward target CEFR levels.

The corpus, the annotation tools, and the models will be freely available via a GitHub repository. Moreover, the tool for automatic text difficulty prediction will be available via a dedicated API.

8. Acknowledgements

This research was supported by the National Recovery and Resilience Plan (NRRP) “Greece 2000” under the “Clusters of Research Excellence” (CREs) program, SUB1.1, with project code $\mathrm{O\Pi\Sigma\ TA}\ 5180519$ and title “Interactive Agent with Emotional Intelligence for Second/Foreign Language Learning”, Acronym: “EmoBot”.

9. Ethical considerations and limitations

Despite its extensibility, the current implementation remains limited to German, with Greek currently under development. Therefore, cross-linguistic generalization has not yet been empirically validated. The corpus used in this study consists exclusively of publicly available written materials, including open-access media sources and sample examination materials released by officially recognized certification bodies. No personal data or learner-produced texts were collected.

10. Bibliographical References

References

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Richard Bamberger and Erich Vanecek. 1984. *Lesen - Verstehen - Lernen - Schreiben. Die Schwierigkeitsstufen von Texten in deutscher Sprache.* Jugend u. Volk Sauerlaender, Wien.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German’s next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Council of Europe, Strasbourg.

Edgar Dale and Jeanne S. Chall. 1949. The concept of readability. *Elementary English*, 26(1):19–26.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17. Association for Computational Linguistics.

R Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1063–1080, Mumbai, India.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Joseph Marvin Imperial. 2021. BERT embeddings for automatic readability assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 611–618.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, et al. 2025. UniversalCEFR: Enabling open multilingual research on language proficiency assessment. *arXiv preprint arXiv:2506.01419*.

Athanasios Kallipolitis, Dionysios Koulouris, Melina Tziokama, Kosmas Pinitas, Argyrios Zafeiriou, Andreas Menychtas, Ilias Maglogiannis, Voula Giouli, Athina Sioupi, Stamatia Michalopoulou, George Tsoulouhas, Michail Katras, Panagiotis Charalampopoulos, and Aristotelis Stamopoulos. 2026. Conversational agent with emotional intelligence for foreign language learning. In *Proceedings of the 14th International Conference on Information and Education Technology (IEEE-ICIET 2026), Koriyama, Japan*. IEEE.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report Research Branch Report 8-75, Naval Technical Training Command Millington TN Research Branch.

David Kogan, Max Schumacher, Sam Nguyen, Masanori Suzuki, Melissa Smith, Chloe Sophia Bellows, and Jared Bernstein. 2025. Ace-CEFR – a dataset for automated evaluation of the linguistic difficulty of conversational texts for LLM applications. *arXiv preprint arXiv:2506.14046*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*. ArXiv:1711.05101.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2016. A readable read: Automatic assessment of language learning materials based on linguistic complexity. *International Journal of Computational Linguistics and Applications*, 7(1):143–159.

Donya Rooein, Paul Röttger, Anastassia Shaitarova, and Dirk Hovy. 2024. Beyond flesch-kincaid: Prompt-based metrics improve difficulty classification of educational texts. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 54–67, Mexico City, Mexico. Association for Computational Linguistics.

Brucce Neves dos Santos, Ricardo Marcondes Marcacini, and Solange Oliveira Rezende. 2021. Multi-domain aspect extraction using bidirectional encoder representations from transformers. *IEEE Access*, 9:91604–91613.

E. A. Smith and R. Senter. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories*, pages 1–14.

Sean Trott and Pamela Rivière. 2024. Measuring and modifying the readability of English texts with GPT-4. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 126–134, Miami, Florida, USA. Association for Computational Linguistics.

Satoru Uchida. 2024. Evaluating the accuracy of ChatGPT in assessing writing and speaking: A verification study using ICNALE GRA. *Learner Corpus Studies in Asia and the World*, 8.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173. Association for Computational Linguistics.

Wiktionary contributors. 2024. Wiktionary, the free dictionary. Accessed: 2025.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22. Association for Computational Linguistics.

Kevin P. Yancey, Geoffrey T. LaFlair, Anthony R. Verardi, and Jill Burstein. 2023. Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584. Association for Computational Linguistics.

11. Appendix

11.1. The sets of features used in Machine Learning

1. **Basic text statistics** (5 features): word count, sentence count, average sentence length, lexical diversity (type–token ratio), and Flesch Reading Ease score.

2. **Readability metrics** (8 features): Wiener Sachtextformel, Flesch–Kincaid Grade Level, Gunning Fog, SMOG Index, ARI, Coleman–Liau, average syllables per word, and average characters per word.

3. **CEFR vocabulary profile** (7 features): percentage of tokens at each CEFR level (A1–C2) and percentage of unknown tokens.

4. **Derived vocabulary ratios** (3 features): basic (A1+A2), intermediate (B1+B2), and advanced (C1+C2) vocabulary proportions.

5. **Morphosyntactic features** (4 features): percentages of irregular verbs, separable verbs, compound words, and international/borrowed words, derived from the Wiktionary-enriched lexicon.

6. **Syntactic complexity** (12 features): subordinate, compound, relative, and infinitive clause ratios; passive voice and modal verb ratios; nominalization ratio; average and maximum clause depth; mean dependency distance; noun–verb ratio; and clauses per sentence.

7. **TF-IDF features** (50 features): unigram and bigram TF-IDF vectors (max 500 terms, sublinear TF scaling, minimum document frequency of 2) reduced to 50 principal components via PCA.

11.2. Confusion Matrices

	A1	A2	B1	B2	C1	C2
A1	**10**	7	2	0	0	0
A2	2	**14**	6	0	3	0
B1	0	3	**18**	4	0	0
B2	0	1	10	**12**	0	0
C1	0	0	3	21	**4**	0
C2	0	0	2	5	7	**0**

Table 7: Confusion matrix for the rule-based classifier. Rows represent true labels; columns represent predicted labels. Note the strong downward bias: 21 of 28 C1 texts are misclassified as B2, and all C2 texts are assigned to lower levels.

	A1	A2	B1	B2	C1	C2
A1	**13**	6	0	0	0	0
A2	4	**20**	1	0	0	0
B1	1	4	**11**	8	1	0
B2	0	1	5	**12**	4	1
C1	0	0	2	1	**20**	5
C2	0	0	0	0	8	**6**

Table 8: Confusion matrix for the feature-based ML classifier (Random Forest). Errors concentrate along the diagonal; the B1 level shows the widest spread, with 8 texts misclassified as B2 and 4 as A2.

	A1	A2	B1	B2	C1	C2
A1	**20**	1	1	0	0	0
A2	2	**22**	1	0	0	0
B1	3	4	**14**	3	0	0
B2	0	1	6	**12**	3	1
C1	0	0	0	7	**15**	5
C2	0	0	0	0	7	**8**

Table 9: Confusion matrix for the fine-tuned BERT model. Tightest diagonal concentration among the three approaches; the largest off-diagonal entries are C1→B2 (7) and C2→C1 (7), both involving adjacent levels.