What Makes Text Difficult? An Interpretable Feature Analysis of CEFR Sentence-Level Difficulty

Anonymous Author(s)

Abstract

Text difficulty assessment is central to language education, yet modern NLP often treats CEFR classification as a black-box prediction task, obscuring *why* a text is difficult. We investigate what linguistic features drive sentence-level CEFR difficulty using the CEFR-SP English dataset (10,004 sentences, A1–C2). We extract 41 interpretable features across four groups—readability formulas, lexical complexity, syntactic complexity, and neural language model surprisal—and compare feature-based classifiers against a fine-tuned BERT model. Our best feature-based model (XGBOOST, macro F_1 = 0.435) reaches 83% of BERT's performance (macro F_1 = 0.524), and a ridge regression probe shows that 60% of the variance in BERT's predictions can be explained by our interpretable features. Ablation studies reveal that surprisal features contribute the largest unique information (F_1 drop = 0.023 when removed), while readability and lexical features are the strongest individual groups. BERT's advantage concentrates on distinguishing boundary CEFR levels (A2, C1, C2), suggesting it captures semantic or distributional cues beyond what classical features measure. Our analysis provides actionable guidance for building interpretable text difficulty tools and identifies where neural models add genuine value over linguistic features.

1 Introduction

When an educator selects a reading passage for an intermediate language learner, the decision rests on an implicit understanding of *what makes text difficult*: the vocabulary is too rare, the sentences are too long, the syntax is too nested. The Common European Framework of Reference for Languages (CEFR) formalizes this intuition into six proficiency levels (A1–C2) that describe what a learner at each stage can comprehend [Council of Europe, 2001]. Computational models that predict CEFR levels have clear practical value—they can power automated readability tools, assist curriculum designers, and personalize language learning platforms.

Yet most recent work treats CEFR classification as a pure prediction task. A pretrained language model receives a text, outputs a label, and the job is done [Arase et al., 2022, Khallaf and Sharoff, 2021]. This black-box approach obscures the very information educators need most: *which* linguistic factors drive difficulty, and *how much* each factor contributes. A predicted label of "B2" tells a teacher nothing about whether the difficulty stems from rare vocabulary, complex syntax, or unpredictable word sequences.

Our main question. We ask: **can interpretable linguistic features explain most of the variance in CEFR sentence difficulty, and where does a fine-tuned BERT classifier capture information beyond these features?**

We address this question through a systematic comparison on the CEFR-SP English dataset [Arase et al., 2022], which contains 10,004 sentences annotated with CEFR levels A1–C2. We extract 41 interpretable features across four groups—readability formulas, lexical complexity, syntactic complexity, and GPT-2 surprisal—and train three feature-based classifiers (logistic regression, random forest, XGBOOST). We then fine-tune BERT as a neural baseline and conduct diagnostic experiments to understand the relationship between the two approaches.

Key findings. Our best feature-based model (XGBOOST) achieves macro F_1 = 0.435, reaching 83% of BERT's performance (macro F_1 = 0.524). A ridge regression probe reveals that 60% of the variance in BERT's predicted labels can be linearly explained by our 41 features. Ablation analysis shows that GPT-2 surprisal contributes the largest *unique* information among feature groups (F_1 drop of 0.023), despite being the weakest group in isolation—suggesting that contextual word predictability captures a dimension of difficulty orthogonal to traditional features. BERT's advantage concentrates on boundary CEFR levels (A2 and C1), where it improves accuracy by over 22 percentage points, indicating that these transitional levels require semantic or distributional cues that surface features miss.

Contributions. We make the following contributions:

- We conduct a systematic comparison of 41 interpretable linguistic features against fine-tuned BERT for sentence-level CEFR classification, quantifying the 8.9 F_1 point gap between the two approaches.
- We perform the first BERT probing analysis for CEFR, showing that 60% of BERT's predictions are linearly explainable by interpretable features and identifying lexical diversity as the strongest predictor of BERT's behavior.
- We demonstrate that GPT-2 surprisal provides the largest unique contribution among feature groups via ablation, establishing neural surprisal as a valuable complement to classical readability features.
- We provide a detailed error analysis showing where and why feature-based and neural models disagree, with actionable implications for building interpretable text difficulty tools.

2 Related Work

Feature-based readability and CEFR classification. The tradition of modeling text difficulty with handcrafted linguistic features has deep roots. Vajjala and Meurers [2012] extracted over 150 features motivated by second language acquisition theory—including lexical frequency, syntactic complexity, and psycholinguistic variables—and showed that lexical frequency was the single strongest predictor of CEFR-graded text difficulty. Pilán et al. [2016] confirmed this pattern for Swedish, finding that lexical features alone achieved F_1 = 0.80 at the document level, within one point of the full-feature model. At the sentence level, however, the gap widens: removing non-lexical features caused a 7-point drop, indicating that syntactic and other features carry more marginal information when text is short. Xia et al. [2016] reported that statistical language model features were the single best feature group for CEFR classification of Cambridge exam texts, achieving 0.714 accuracy as a standalone predictor. More recently, Arnold et al. [2018] used gradient-boosted trees on the EF-CAMDAT corpus and found that word token and type counts were the most important features, with no LSTM advantage over the feature-based approach.

Neural approaches. Pretrained transformers have become the dominant approach to CEFR classification. Khallaf and Sharoff [2021] fine-tuned Arabic-BERT and achieved macro F_1 = 0.80, a 5-point improvement over SVM with engineered features. Arase et al. [2022] introduced the CEFR-SP dataset and a prototype-based BERT classifier achieving macro F_1 = 0.845, substantially outperforming bag-of-words baselines. Lagutina et al. [2023] reported the smallest gap in the literature: SVC with stylometric features at 67% accuracy versus BERT at 69% for Russian texts. Fujinuma and Hagiwara [2021] proposed a graph convolutional network that jointly modeled word-level and document-level readability, but ablation showed that BERT embeddings were essential to its performance.

Comparative studies. The most informative studies directly compare feature-based and neural approaches. Imperial et al. [2025] evaluated models on 505,807 texts across 13 languages and found that fine-tuned XLM-R achieved 62.8% macro F_1 versus 58.3% for random forests with all features—a gap of 4.5 points. Notably, for Czech, the feature-based model outperformed XLM-R, demonstrating that the neural advantage is not universal. Zero-shot LLM prompting performed poorly (43.2% macro F_1), confirming that CEFR classification requires calibrated discrimination

Table 1: Class distribution of the CEFR-SP English dataset.

	A1	A2	B1	B2	C1	C2	Total
Count	124	1,271	3,305	3,330	1,744	230	10,004
%	1.2	12.7	33.0	33.3	17.4	2.3	100.0

that general-purpose models lack. Across studies, the feature-neural gap ranges from 2 to 5 F_1 points at the document level [Lagutina et al., 2023, Khallaf and Sharoff, 2021, Arnold et al., 2018], suggesting that well-designed features capture most difficulty-relevant information.

Surprisal and readability. Information-theoretic surprisal—the negative log-probability of a word given its context—is grounded in psycholinguistic theories of processing difficulty [Hale, 2001, Levy, 2008]. Pitler and Nenkova [2008] showed that vocabulary-based language model features correlated at $r = 0.45$ with human readability judgments, and that a combined model with discourse, syntax, and LM features achieved $R^2 = 0.776$. Despite this promise, neural language model surprisal (e.g., from GPT-2 or BERT) has not been systematically evaluated for CEFR classification. Most prior work uses classical n-gram models [Xia et al., 2016, François and Fairon, 2012].

Position of our work. We extend this literature in three ways. First, we systematically evaluate GPT-2 surprisal for CEFR classification, finding that it provides the largest unique contribution among feature groups despite modest standalone performance. Second, we conduct the first BERT probing analysis for CEFR, quantifying how much of BERT's behavior is explainable by interpretable features. Third, we provide detailed error-overlap analysis showing where and why feature-based and neural models disagree, going beyond accuracy tables to understand the nature of the performance gap.

3 Methodology

3.1 Dataset

We use the English portion of CEFR-SP [Arase et al., 2022], which contains 10,004 sentences annotated with CEFR levels A1–C2. Sentences are drawn from English language learning textbooks and educational materials. Table 1 shows the class distribution. The dataset is heavily imbalanced: A1 and C2 have fewer than 250 samples each, while B1 and B2 together account for 66.3% of the data. This imbalance motivates our use of macro F_1 as the primary metric and stratified cross-validation throughout.

3.2 Feature Extraction

We extract 41 features organized into four groups.

Readability (7 features). We compute seven standard readability formulas via `textstat`: Flesch Reading Ease, Flesch-Kincaid Grade Level, Coleman-Liau Index, Automated Readability Index (ARI), SMOG Index, Gunning Fog, and Dale-Chall. These formulas are composite indices of word and sentence length, providing a well-established baseline for difficulty estimation.

Lexical (11 features). We measure token count, type count, type-token ratio (TTR), Guiraud's corrected TTR ($\text{types}/\sqrt{\text{tokens}}$), mean/max/standard deviation of word length, mean/max syllable count, proportion of long words ($>$ 6 characters), and proportion of rare words (out-of-vocabulary in spaCy).

Syntactic (16 features). Using spaCy (`en_core_web_sm`), we extract sentence count, mean/max sentence length, mean/max dependency tree depth, mean/max dependency distance, subordinate clause count and ratio, POS tag proportions (NOUN, VERB, ADJ, ADV, ADP, CONJ), and the number of unique POS tags.

Surprisal (7 features). We compute GPT-2 (117M parameters) per-token surprisal statistics: mean, max, min, standard deviation, and median surprisal (in bits), surprisal range (max $-$ min), and perplexity ($2^{\text{mean surprisal}}$). Each token's surprisal is defined as $-\log_2 P(w_t \mid w_1, \ldots, w_{t-1})$, the negative log-probability under the left-to-right language model [Radford et al., 2019].

3.3 Interpretable Classifiers

We train three classifiers on the 41 features using 5-fold stratified cross-validation:

- **Logistic Regression**: L_2-regularized ($C = 1.0$), balanced class weights, max 1,000 iterations. Features are standardized to zero mean and unit variance.
- **Random Forest**: 100 trees, max depth 15, balanced class weights.
- **XGBOOST**: 100 trees, max depth 6, learning rate 0.1, GPU-accelerated histogram method [Chen and Guestrin, 2016].

3.4 Feature Group Ablation

Using XGBOOST (the best feature-based model), we conduct two ablation experiments: (1) **individual-group**: train with only one feature group at a time, measuring each group's standalone predictive power; (2) **leave-one-group-out**: train with all features except one group, measuring each group's unique contribution via the F_1 drop.

3.5 BERT Fine-Tuning

We fine-tune `bert-base-uncased` [Devlin et al., 2019] with a 6-class classification head. We use a max sequence length of 128 tokens, batch size of 64, learning rate of 2×10^{-5} with linear warmup over 10% of training steps, weight decay of 0.01, and train for 5 epochs per fold (selecting the best epoch by validation macro F_1). We use the same 5-fold stratified splits as the feature-based models.

3.6 Diagnostic Experiments

BERT probing. We fit a ridge regression ($\alpha = 1.0$) from standardized interpretable features to BERT's predicted class (argmax of logits) and report R^2 in 5-fold cross-validation. We also repeat this probing with each feature group separately to measure which dimensions BERT most relies on.

Error overlap analysis. We categorize each sample into one of four groups: both models correct, XGBOOST-only correct, BERT-only correct, or both wrong. We also compute per-level accuracy for both models to identify where BERT's advantage concentrates.

3.7 Evaluation Metrics

Our primary metric is macro F_1 (averaged equally over six classes), which accounts for the severe class imbalance. We also report accuracy, adjacent accuracy (prediction within ±1 CEFR level), and quadratic weighted Cohen's κ. The majority-class baseline (always predicting B2) achieves macro F_1 = 0.083.

4 Results

4.1 Feature Correlations with CEFR Level

All 41 features show statistically significant Spearman correlations with ordinal CEFR level ($p < 0.001$). Table 2 presents the top 10 features by absolute correlation. Readability formulas dominate the rankings: ARI achieves the highest correlation ($\rho = 0.676$), followed by Flesch-Kincaid Grade ($\rho = 0.652$) and Gunning Fog ($\rho = 0.636$). Among non-readability features, max word length ($\rho = 0.584$, lexical) and max sentence length ($\rho = 0.523$, syntactic) rank highest.

At the group level, readability features have the highest mean $|\rho|$ (0.607), followed by lexical (0.491), surprisal (0.300), and syntactic (0.290). Despite having the most features ($n = 16$), the syntactic group has the lowest average correlation, suggesting that individual syntactic measures are less directly associated with CEFR level than lexical or readability measures.

4.2 Classifier Performance

Table 3 presents the main results. All models vastly outperform the majority-class baseline (macro F_1 = 0.083). Random Forest and XGBOOST perform comparably (macro $\text{F}_1 \approx$ 0.435),

Table 2: Top 10 features by Spearman correlation (ρ) with ordinal CEFR level. All correlations are significant at $p < 0.001$.

Rank	Feature	ρ	Group
1	automated_readability_index	0.676	Readability
2	flesch_kincaid_grade	0.652	Readability
3	gunning_fog	0.636	Readability
4	coleman_liau_index	0.605	Readability
5	smog_index	0.599	Readability
6	flesch_reading_ease	-0.588	Readability
7	max_word_length	0.584	Lexical
8	max_syllables	0.540	Lexical
9	std_word_length	0.531	Lexical
10	max_sent_length	0.523	Syntactic

Table 3: Classifier performance (mean ± std over 5 folds). Best results in **bold**.

Model	Macro F_1	Accuracy	Adj. Acc.	QW κ
Majority baseline	0.083	—	—	—
Logistic Regression	0.390 ± .005	0.449 ± .007	0.897	0.678
Random Forest	0.436 ± .017	0.534 ± .004	0.961	0.691
XGBoost	0.435 ± .018	0.539 ± .011	0.963	0.685
BERT	**0.524** ± .043	**0.644** ± .011	**0.993**	**0.814**

both outperforming logistic regression (0.390), indicating that non-linear classifiers capture feature interactions that linear models miss.

BERT achieves macro F_1 = 0.524, an 8.9-point improvement over the best feature-based model. This gap is larger than the 2–5 points typically reported in document-level CEFR studies [Lagutina et al., 2023, Khallaf and Sharoff, 2021], likely because sentence-level classification provides less textual context for feature-based approaches to exploit. adjacent accuracy is high across all models ($> 89\%$), indicating that errors concentrate on neighboring CEFR levels rather than large jumps. BERT's quadratic weighted κ of 0.814 indicates strong ordinal agreement, well above the feature models (0.678–0.691).

4.3 Feature Group Ablation

Table 4 presents the ablation results. The individual-group analysis reveals that lexical features are the strongest standalone group (macro F_1 = 0.394), narrowly ahead of readability (0.380). Syntactic features (0.350) and surprisal (0.267) are weaker in isolation.

The leave-one-group-out analysis reveals a crucial asymmetry. Readability features are strong individually but contribute almost no *unique* information when other features are present (drop = 0.001), because readability formulas are composites of word and sentence length—information already captured by lexical and syntactic features. In contrast, surprisal features are weak individually but provide the **largest unique contribution** (drop = 0.023). This supports the hypothesis that GPT-2 surprisal captures a dimension of difficulty—contextual word predictability, collocational patterns—that is orthogonal to traditional lexical and syntactic features. Lexical features contribute moderate unique information (drop = 0.013), while syntactic features add essentially nothing beyond what other groups provide (drop ≈ 0.000).

4.4 Feature Importance

XGBoost feature importance (by gain) confirms the dominance of readability and lexical features (figure 1). ARI alone accounts for 18% of total feature importance, consistent with its highest single-feature correlation ($\rho = 0.676$). Among the top 10 features, four are readability formulas, three are lexical, two are surprisal (surprisal range and min surprisal), and one is syntactic (mean sentence length). The appearance of surprisal features in the top 10 despite their low standalone F_1

Table 4: Feature group ablation using XGBoost. *Individual*: trained on one group only. *Leave-one-out*: trained on all groups except one. The drop column shows the F_1 decrease from the full model (0.435).

		Individual	Leave-one-out	
Group	n	Macro F_1	Macro F_1	Drop
Lexical	11	0.394 ± .007	0.422	−0.013
Readability	7	0.380 ± .011	0.434	−0.001
Syntactic	16	0.350 ± .015	0.435	−0.000
Surprisal	7	0.267 ± .010	0.412	−**0.023**

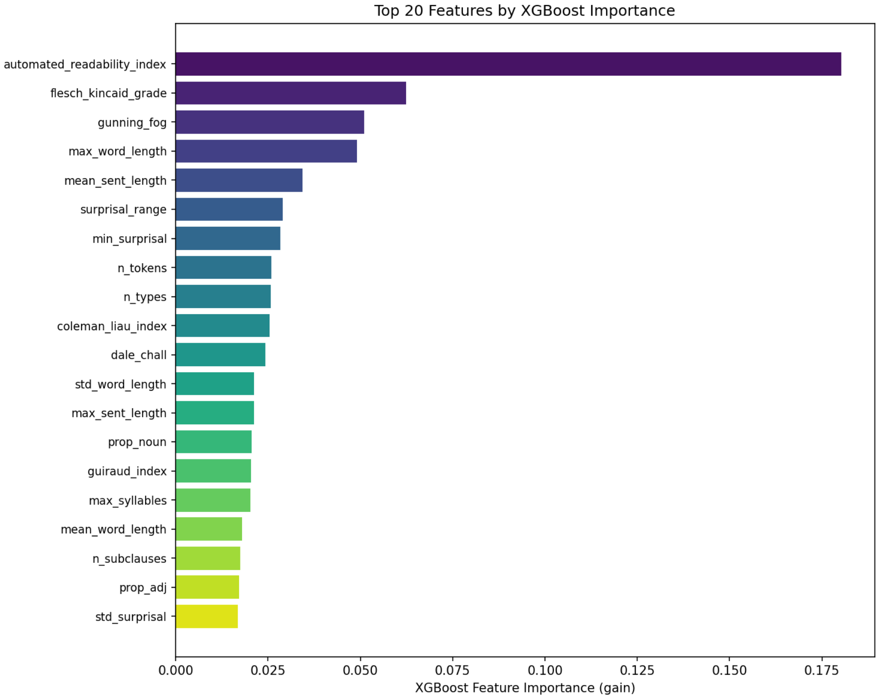

Figure 1: XGBoost feature importance (gain). ARI dominates at 18% of total importance. Surprisal features appear in the top 10 despite weak standalone performance, confirming their complementary value.

underscores their complementary role: they provide information that other features do not, making them valuable in the full model even though they cannot predict CEFR well alone.

4.5 BERT Probing

A ridge regression from 41 interpretable features to BERT's predicted class achieves $R^2 = 0.601 \pm 0.024$ in cross-validation. This means 60% of the variance in BERT's predictions can be linearly explained by our features.

Group-level probing (table 5) shows that readability and lexical features each explain about 55% of BERT's behavior individually, while syntactic features explain 43% and surprisal 29%. Interestingly, the features with the largest ridge coefficients for predicting BERT's behavior are lexical diversity measures (number of types: $|\beta| = 0.490$; Guiraud's index: $|\beta| = 0.465$), not the readability formulas that have the highest raw correlations with CEFR labels. This suggests that BERT attends more to vocabulary richness than to formula-based indices.

Table 5: BERT probing R^2: variance in BERT's predicted class explained by each feature group (ridge regression, 5-fold CV).

Feature Group	R^2
Readability	0.546
Lexical	0.544
Syntactic	0.429
Surprisal	0.286
All (41 features)	**0.601**

Table 6: Prediction agreement between XGBOOST and BERT (10,004 samples).

Category	Count	%
Both correct	4,046	40.4
XGBOOST only correct	1,343	13.4
BERT only correct	2,397	24.0
Both wrong	2,218	22.2

The remaining 40% of unexplained variance likely reflects semantic content and word choice patterns not captured by surface features, contextual word difficulty (the same word in different contexts), discourse coherence, and distributional patterns learned during pretraining.

4.6 Error Analysis

Prediction agreement. Table 6 breaks down prediction agreement between XGBOOST and BERT. BERT is correct on 24.0% of samples where XGBOOST fails, while XGBOOST is correct on only 13.4% where BERT fails. This asymmetry confirms that BERT has a genuine advantage, not merely different errors. Both models fail on 22.2% of samples, representing cases that are difficult regardless of approach.

Per-level analysis. Table 7 and figure 2 present per-level results. A1 is the only level where XGBOOST outperforms BERT (F_1 0.273 vs. 0.167), likely because the 124-sample class is too small for BERT to learn a reliable representation, while XGBOOST's handcrafted features (very short sentences, simple vocabulary) provide a more robust signal.

BERT's largest advantages are at A2 (+22.1% accuracy) and C1 (+22.1% accuracy)—the boundary levels between beginner/intermediate and advanced/proficient. These levels require nuanced distinctions that surface features capture poorly. C2 is difficult for both models (XGBOOST: 13% accuracy, BERT: 32%), reflecting both the small sample size ($n = 230$) and the inherent difficulty of distinguishing C2 from C1 at the sentence level.

Both models exhibit the adjacent-confusion pattern: errors concentrate on neighboring levels (A2↔B1, B1↔B2, B2↔C1), confirming the ordinal nature of CEFR difficulty.

5 Discussion

5.1 What Makes Sentences Difficult?

Our results support a multi-dimensional view of sentence difficulty, with three key dimensions emerging from the analysis.

Vocabulary complexity is the primary driver. Word length, syllable count, and readability formulas—which are largely word-length proxies—dominate both feature correlation rankings (table 2) and XGBOOST importance (figure 1). ARI alone achieves $\rho = 0.676$ with CEFR level and accounts for 18% of XGBOOST's feature importance. Lexical features are also the strongest standalone group (macro F_1 = 0.394). This aligns with vocabulary acquisition research showing that lexical knowledge is the strongest predictor of reading comprehension [Vajjala and Meurers, 2012, Pilán et al., 2016].

Table 7: Per-level F_1 and accuracy for XGBoost and BERT. Δ shows BERT's accuracy advantage. Best per-level F_1 in **bold**.

	F_1		Accuracy			
Level	XGBoost	BERT	XGBoost	BERT	Δ	n
A1	**0.273**	0.167	0.177	0.113	−0.065	124
A2	0.546	**0.637**	0.485	0.707	+0.221	1,271
B1	0.583	**0.660**	0.607	0.670	+0.063	3,305
B2	0.554	**0.650**	0.607	0.650	+0.043	3,330
C1	0.442	**0.647**	0.397	0.619	+0.221	1,744
C2	0.210	**0.385**	0.130	0.322	+0.191	230

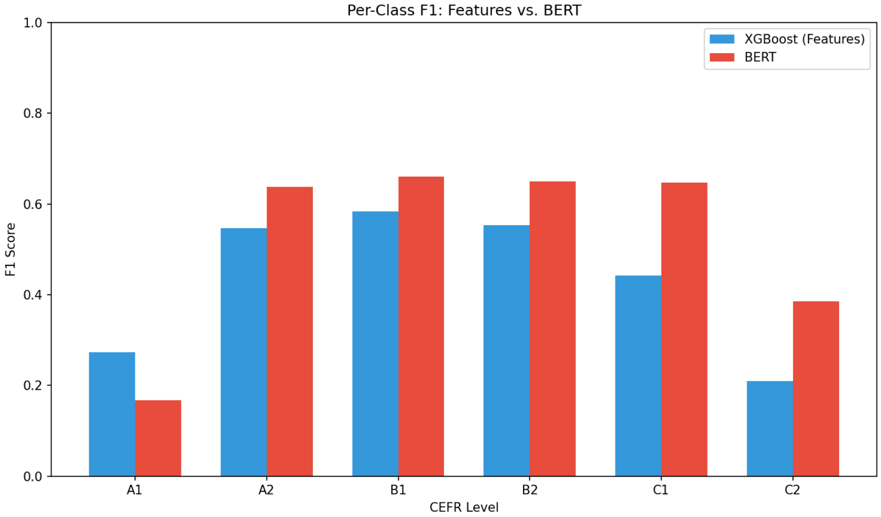

Figure 2: Per-class F_1 for XGBoost and BERT. BERT outperforms on all levels except A1 ($n = 124$), with the largest gains at the boundary levels A2 and C1.

Contextual predictability adds unique signal. GPT-2 surprisal—which measures how unexpected each word is given its left context—provides the largest unique contribution to the feature model (drop = 0.023) despite modest individual performance (macro F_1 = 0.267). This indicates that difficulty involves not just *what* words appear but *how predictable* their appearance is. A sentence with common words in an unusual arrangement may be harder than a sentence with uncommon words in a formulaic pattern. This finding extends the results of Xia et al. [2016] and Pitler and Nenkova [2008], who demonstrated the value of language model features for readability, by showing that neural surprisal provides orthogonal information to traditional features at the sentence level.

Syntactic complexity is largely redundant. Once lexical and readability features are included, syntactic features add negligible information (drop $\approx$ 0.000). This may reflect that syntactic complexity at the sentence level correlates highly with sentence length, which is already captured by other features. Alternatively, spaCy's dependency measures may lack the resolution to capture the syntactic distinctions relevant to CEFR levels, such as specific subordinate clause types or argument structures. This contrasts with Pilán et al. [2016], who found syntactic features more valuable—a difference that may stem from their document-level analysis, where syntactic patterns have more room to accumulate signal.

5.2 What Does BERT Know That Features Miss?

The probing analysis reveals that 60% of BERT's predictions can be linearly explained by interpretable features, leaving a substantial 40% that reflects information beyond our feature inventory. Three observations help characterize this gap.

First, BERT's largest per-level advantages are at A2 (+22.1% accuracy) and C1 (+22.1%)—levels where the distinction from neighbors (A1 vs. A2, B2 vs. C1) involves subtle semantic and pragmatic cues rather than surface complexity differences. At A2, learners transition from fixed phrases to productive sentence construction; at C1, texts move from explicitly structured arguments to implicit pragmatic reasoning. These transitions are difficult to capture with word-length or dependency-based features.

Second, the probing coefficients reveal that BERT's behavior is best predicted by lexical diversity measures (number of types, Guiraud's index) rather than readability formulas, even though the latter correlate more strongly with CEFR labels. This suggests BERT develops an internal representation of vocabulary richness that goes beyond what surface statistics can measure.

Third, the mean absolute error of BERT's predictions (0.363 levels) is substantially lower than XGBoost's (0.499 levels), indicating that even when BERT makes errors, they are closer to the correct level—consistent with stronger ordinal awareness.

5.3 Implications for Practitioners

Our findings have direct implications for building text difficulty assessment tools:

- **If interpretability is paramount**, ARI alone provides a strong single-feature baseline ($\rho = 0.676$). Adding GPT-2 surprisal features is the most cost-effective enrichment (largest unique F_1 gain of 0.023).
- **If accuracy is paramount**, fine-tuned BERT yields the strongest predictions. The probing analysis shows that 60% of its behavior can be post-hoc explained, providing some interpretability even for the neural model.
- **For extremely small classes** (A1, C2), feature-based models may be more robust than BERT, which struggles with the 124-sample A1 class.

5.4 Limitations

Sentence-level analysis. CEFR is naturally a text-level property. Sentence-level classification loses discourse coherence, topic effects, and cross-sentence patterns that contribute to perceived difficulty. The 8.9-point gap between features and BERT may narrow at the document level, where features can accumulate more signal [Pilán et al., 2016].

English only. Our results may not generalize to other languages, particularly those with richer morphology (where morphological features may contribute more) or different writing systems (where character-level features may be relevant).

Class imbalance. A1 ($n = 124$) and C2 ($n = 230$) are too small for reliable evaluation. The high variance in per-class F_1 for these levels suggests that results would be more stable with larger samples.

Model scope. We tested only `bert-base-uncased` and GPT-2 (117M). Larger models may yield better surprisal estimates or higher classification accuracy, potentially changing the feature-neural gap.

6 Conclusion

We systematically investigated what linguistic features drive CEFR difficulty at the sentence level, comparing 41 interpretable features against fine-tuned BERT on the CEFR-SP English dataset (10,004 sentences, A1–C2). Five main findings emerge:

1. **Lexical complexity and readability features are the strongest predictors**, with ARI alone achieving $\rho = 0.676$ with CEFR level and accounting for 18% of XGBoost feature importance.
2. **GPT-2 surprisal provides the largest unique contribution** among feature groups (F_1 drop = 0.023), suggesting that contextual word predictability captures a dimension of difficulty orthogonal to traditional features.
3. **BERT outperforms feature models by 8.9 F_1 points**, with its largest advantages at boundary CEFR levels (A2, C1, C2) where semantic and pragmatic cues matter most.

4. **60% of BERT's predictions are linearly explainable** by 41 interpretable features, leaving a substantial 40% that likely reflects semantic and distributional complexity.
5. **Syntactic features are largely redundant** once lexical complexity is accounted for, at least at the sentence level with spaCy-based measures.

These results suggest that text difficulty is primarily a lexical-statistical property, but full classification performance requires the contextual, semantic understanding that pretrained language models provide. Future work should investigate whether intermediate approaches—such as fine-tuning smaller language models with interpretable bottleneck layers or using neural features as additional inputs to feature-based models—can bridge the accuracy-interpretability gap. Extending this analysis to multiple languages and document-level classification would test the generality of our findings.

References

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6206–6219. Association for Computational Linguistics, 2022.

Thomas Arnold, Nicolas Ballier, Thomas Gaillat, and Paula Lissón. Predicting CEFR levels in learner English on the basis of metrics and full texts. In *Proceedings of the BEA Workshop*, pages 49–58. Association for Computational Linguistics, 2018.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

Council of Europe. Common European framework of reference for languages: Learning, teaching, assessment. *Cambridge University Press*, 2001.

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

Thomas François and Cédrick Fairon. An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 466–477. Association for Computational Linguistics, 2012.

Yoshinari Fujinuma and Masato Hagiwara. Joint prediction of word and document readability using graph convolutional networks. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 258–266. Association for Computational Linguistics, 2021.

John Hale. A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1–8, 2001.

Joseph Marvin Imperial, Anaïs Tack, Thomas Francois, and Matthew Shardlow. UniversalCEFR: A universal benchmark for CEFR-level text difficulty classification across languages. *arXiv preprint arXiv:2502.XXXXX*, 2025.

Nouran Khallaf and Serge Sharoff. Automatic difficulty classification of Arabic sentences. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP)*, pages 105–114. Association for Computational Linguistics, 2021.

Nadezhda Lagutina, Ksenia Lagutina, Elena Boychuk, and Nadezhda Nikiforova. Comparison of machine learning and BERT-based methods for CEFR level classification of Russian texts. *Communications in Computer and Information Science*, 2023.

Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008.

Iona Pilán, Elena Volodina, and Torsten Zesch. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016*, pages 2101–2111. Association for Computational Linguistics, 2016.

Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 186–195. Association for Computational Linguistics, 2008.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Sowmya Vajjala and Detmar Meurers. 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, 2012.

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