SAAKTH at BEA 2026 Shared Task 1: L1-Aware English Vocabulary Difficulty Prediction with Hybrid Transformer and Psycholinguistic Features

Karthik Mattu **Adit Dhall** **Arshad Naguru**
Shubh Sehgal **Thejas N. Gowda** **Hakyung Sung**
Rochester Institute of Technology
{km5503, ad6449, an2629, ss8179, tl6153, hksgla}@rit.edu

Abstract

This paper presents team SAAKTH's system for the BEA 2026 Shared Task on Vocabulary Difficulty Prediction (Closed Track). We address the key challenge that English word difficulty is not fixed but varies with English learners' native language. Our approach combines a fine-tuned XLM-RoBERTa-large encoder with handcrafted psycholinguistic features engineered separately for each L1 group. These features are integrated via a shallow multilayer perceptron and optimized separately per L1, with five-seed ensembling and XGBoost-based blending for stability. Our system achieves RMSEs of 0.997 (es), 1.002 (de), and 0.932 (cn) on the development set, improving 20–25% over the baseline. Results highlight the effectiveness of L1-aware modeling under limited data.

1 Introduction

Predicting English word difficulty for second language (L2) learners has clear value for language pedagogy, adaptive testing, and content recommendation (Settles et al., 2020). A key challenge, however, is that difficulty is not a fixed property of words; it depends on the learner's first language (L1) background (Ringbom, 2007). For instance, an English learner whose L1 is Spanish may readily recognize *university* through its cognate *universidad*, whereas an English learner whose L1 is Mandarin has no such advantage.

Most prior work on lexical complexity has overlooked this L1-dependence. Shared tasks on Complex Word Identification (Paetzold and Specia, 2016; Yimam et al., 2018), Lexical Complexity Prediction (Shardlow et al., 2021), and Lexical Simplification (Shardlow et al., 2024) have advanced the field but were not designed to account for learner-specific factors such as L1 background despite cross-linguistic influence being central to vocabulary acquisition (Nation, 2001; Odlin, 2003).

The BEA 2026 Shared Task (Felice and Skidmore, 2026) frames English vocabulary difficulty prediction as an L1-aware regression task over three L1 groups: Spanish, German, and Mandarin. Our preliminary analysis of the training data reveals cross-L1 score correlations ranging from $r = 0.63$ to 0.68, meaning approximately 40–46% of difficulty variance is shared across groups. The remaining variance is L1-specific, motivating separate models per language group.

We propose a hybrid system combining XLM-RoBERTa (Conneau et al., 2020) with targeted psycholinguistic features. While transformers capture cross-lingual patterns, they lack explicit access to extralinguistic properties (e.g., age of acquisition, Mandarin stroke complexity). Conversely, handcrafted features are interpretable but fail to capture context-sensitive usage. This tension motivates our hybrid approach. Our main contributions are: (1) A **hybrid architecture** integrating XLM-RoBERTa-large embeddings with psycholinguistic features via a dual-learning-rate MLP; (2) **L1-specific feature engineering**: cognate distance for Spanish and German, and stroke complexity for Mandarin; (3) A **robust ensembling framework** combining five-seed averaging with XGBoost blending, achieving 20–25% root mean square error (RMSE) reduction over the official baseline. All code, including data processing, feature extraction, modeling, and analysis scripts, is publicly available to support reproducibility and future research: https://github.com/aditdhall/bea2026-vocab-difficulty.

2 Related work

Lexical complexity prediction. Word difficulty prediction has been extensively studied in prior work. Earlier work framed the problem as binary Complex Word Identification, classifying words as either difficult or not (Paetzold and Specia, 2016;

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1010–1015

Yimam et al., 2018). More recent shared tasks (e.g., BEA 2021; Shardlow et al., 2021; BEA 2024; Shardlow et al., 2024) have reformulated it as a regression problem, predicting continuous difficulty scores. This shift enables a more fine-grained and nuanced representation of lexical difficulty.

However, a key limitation of this line of work is that it largely treats word difficulty as universal across learners, without accounting for differences in learners' L1 backgrounds. The BEA 2026 dataset (Skidmore et al., 2025), based on the British Council's Knowledge-based Vocabulary Lists (KVL; Schmitt et al., 2024),[1] directly addresses this gap.

Multilingual transformers. Multilingual transformer models such as XLM-RoBERTa (Conneau et al., 2020), trained on large-scale multilingual corpora, are well suited to this task due to their ability to learn cross-lingual representations.

The official baseline models, provided by the shared task organizers (Felice and Skidmore, 2026),[2] fine-tuned XLM-RoBERTa-*base* separately for each L1 with a linear regression head, achieving RMSE of 1.357 (Spanish [es]), 1.328 (German [de]), and 1.175 (Mandarin [cn]).

Feature-based and L1-aware modeling. A substantial body of research shows that word-level features such as word frequency (Brysbaert and New, 2009), age of acquisition (Kuperman et al., 2012), concreteness (Brysbaert et al., 2014), and Common European Framework of Reference for Languages (CEFR) level (Council of Europe, 2001) are strong predictors of lexical difficulty (Laufer, 1997; Nation, 2001) and are associated with broader aspects of L2 proficiency, including lexical development (Crossley et al., 2016) and writing quality (Kim et al., 2018; Sung et al., 2025).

A limitation of previous approaches is that they relied on general lexical features, which may not capture L1-specific sources of difficulty. Additional features are therefore needed to account for cross-linguistic variation: for learners with European L1s, orthographic similarity to L1 translations tends to play a role (Schepens et al., 2012; Dijkstra et al., 2010), whereas for learners with Mandarin L1, character-level properties (e.g., stroke complexity) could be more informative (Shu et al., 2003).

[1] KVL is a multilingual resource with psychometrically calibrated difficulty scores derived from real learner data, providing L1-specific estimates of vocabulary difficulty.

[2] https://github.com/britishcouncil/bea2026st

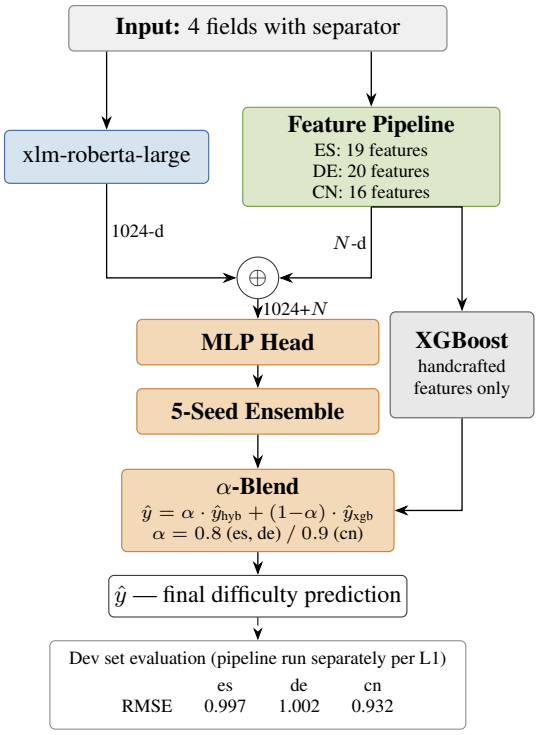

Figure 1: Architecture overview. The encoder produces a 1024-dimensional representation, which is concatenated with an N-dimensional handcrafted feature vector, where $N \in \{19, 20, 16\}$ depending on the L1 (ES/DE/CN). Model predictions are evaluated on the dev set.

3 Method

3.1 Architecture

Our system comprises three main components: a Transformer encoder, a feature extractor, and a fusion multilayer perceptron (MLP). An overview of the full architecture is shown in Figure 1.

Transformer encoder. Building on the organizers' XLM-RoBERTa-base baseline, we adopt the larger XLM-RoBERTa-*large* model to leverage its greater representational capacity for capturing subtle lexical and cross-lingual patterns. This 24-layer model has 1024-dimensional hidden states and is pretrained on 2.5TB of multilingual CommonCrawl data (Conneau et al., 2020). Each input was constructed by concatenating four fields separated by [SEP] tokens: L1_source_word [SEP] L1_context [SEP] en_target_clue [SEP] en_target_word, and processed as a single sequence (Wolf et al., 2020). We used the first-token (<s>) representation (1024-dim) as the item-level embedding.

Feature extractor. We computed a vector of handcrafted features for each item: 19 features for Spanish, 20 for German, and 16 for Mandarin. These are drawn from a fixed set of 21 candidate features (see Section 3.2), where the active subset per L1 was selected once on the training data and kept fixed throughout all experiments.

Fusion MLP. We concatenated the transformer embedding with the feature vector and passed it through a two-layer feed-forward network with ReLU activation and 0.1 dropout to produce a single difficulty score. The model was trained end-to-end using mean square error (MSE) loss.

3.2 Feature engineering

We grouped our features into six categories based on the type of information they capture.

Orthographic features. We included word length, syllable count, the presence of a derivational suffix, and reveal ratio (i.e., the proportion of letters shown in the spelling clue). Reveal ratio proved particularly informative, correlating with GLMM scores at $r \approx 0.35$–0.45 across all L1s.

Frequency and psycholinguistic features. We included SUBTLEX-US log frequency (Brysbaert and New, 2009), age of acquisition (Kuperman et al., 2012), concreteness (Brysbaert et al., 2014), and CEFR proficiency level (Council of Europe, 2001), mapped from the Oxford 5000 word list. Frequency and age of acquisition were the strongest individual predictors ($r = 0.44$–0.62 and $r = 0.43$–0.56, respectively), while CEFR level also showed substantial correlation ($r = 0.47$–0.53).

Semantic and lexical features. We included the number of WordNet synsets (Miller, 1995), maximum hypernym depth, and one-hot part-of-speech indicators (noun, verb, adjective).

Contextual features. We included context sentence length and the position of the target word within the sentence.

Cognate and cross-linguistic features. For Spanish and German, we included features that capture the orthographic similarity between the English word and its L1 translation. This included Levenshtein edit distance (Levenshtein, 1966) ($r = -0.24$ for Spanish, -0.34 for German), character n-gram overlap, a binary cognate flag (distance < 0.4), length ratio, and a German compound word indicator. Together, these features model cross-linguistic similarity (Schepens et al., 2012; Dijkstra et al., 2010).

Stroke count. For Mandarin, we used the total stroke count of the Chinese character(s), obtained from the Unicode *Unihan* database (Unicode Consortium, 2024).[3] This feature was the most predictive feature for Mandarin in our SHAP analysis (mean $|\text{SHAP}| = 0.078$).

We froze the feature inventory early to keep the feature space stable across experiments and avoid repeated tuning on the dev set. This results in 19 features for Spanish (15 shared + 4 cognate), 20 for German (15 shared + 5 cognate), and 16 for Mandarin (15 shared + 1 stroke). The shared feature set comprises 4 orthographic, 4 psycholinguistic, 5 semantic, and 2 contextual features.

4 Experiment

4.1 Setup

Data. Each L1 dataset contained 6,091 training items and 677 development items. Each instance included an English target word, its part-of-speech, a partial spelling clue, the L1 translation, and a context sentence in the learner's language. The target variable was a GLMM-based difficulty score, where lower values indicate higher difficulty.

Implementation. Models were implemented in PyTorch (Paszke et al., 2019) using the HuggingFace Transformers library (Wolf et al., 2020), and trained on an A100 GPU.

4.2 Training

We trained a separate model for each L1 using AdamW (Loshchilov and Hutter, 2019), with learning rates of 1×10^{-5} for the transformer weights and 1×10^{-4} for the MLP head (Howard and Ruder, 2018). Training used a batch size of 16 on an A100 GPU, for up to 8 epochs and patience-3 early stopping based on dev RMSE. Hyperparameters (learning rate, batch size, warmup ratio, and epochs) were identified by evaluating 15 candidate configurations on the Spanish dataset and then applied consistently across all L1 models.

[3] Stroke complexity is known to affect character processing (Shu et al., 2003). We hypothesize that increased complexity may influence L1 processing and acquisition, which may in turn affect the perceived difficulty of corresponding English words for Mandarin learners.

4.3 Ensembling and blending

We found that single-seed runs had noticeable variance, so we trained five models per L1 with seeds [10, 42, 123, 456, 789] and averaged their predictions (Dodge et al., 2020). We further trained an XGBoost model (Chen and Guestrin, 2016) on handcrafted features and linearly blended it with the hybrid ensemble: $\hat{y} = \alpha \cdot \text{hybrid} + (1-\alpha) \cdot \text{xgboost}$. The blend weight α, tuned on the dev set, was 0.8 for Spanish and German and 0.9 for Mandarin, following the idea of stacked generalization (Wolpert, 1992).

4.4 Evaluation

We compared our system against two baselines. The first was the official closed-track baseline provided by the shared task organizers, which fine-tuned XLM-RoBERTa-base for each L1 using a linear regression head. The second was a feature-only XGBoost model trained exclusively on our handcrafted features, with no transformer component. All systems were evaluated on the dev set using RMSE.

5 Results

System	es	de	cn
Official baseline	1.357	1.328	1.175
XGBoost only	1.330	1.310	1.228
Hybrid (5-seed ensemble)	1.021	1.013	0.940
Hybrid + XGBoost blend (dev)	0.997	1.002	0.932
Hybrid + XGBoost blend (test)	**1.045**	**0.994**	**0.900**

Table 1: Dev and test set RMSE. Individual seed RMSE ranged from 1.04–1.16 (es), 1.00–1.10 (de) and 0.96–1.03 (cn). The blend weight α was tuned on the dev set (α=0.8 for es/de; α=0.9 for cn). Test results were obtained using the same ensemble and blend configuration as the dev set.

On the dev set, our best system—the Hybrid + XGBoost blend—achieves RMSEs of 0.997 (es), 1.002 (de), and 0.932 (cn), corresponding to reductions of 26.5%, 24.5%, and 20.7% over the official baseline. The 5-seed ensemble alone reduced RMSE by 24.8% (es), 23.7% (de), and 20.0% (cn). The XGBoost-only model also outperformed the baseline for Spanish (es) and German (de), indicating that the handcrafted features capture meaningful signal.[4]

[4]We did not evaluate a standalone XLM-RoBERTa-large without additional features; therefore, we could not isolate the extent to which performance gains are attributable to the larger model versus the inclusion of handcrafted features.

5.1 Ablation

We conducted ablation experiments on Spanish (single seed) by removing one feature group at a time (Table 2). Frequency and psycholinguistic features contributed the most, with their removal leading to the largest performance degradation ($\Delta = +0.012$). Cognate features improved performance for Spanish ($\Delta = +0.008$), but their removal unexpectedly improved performance for German ($\Delta = -0.034$).[5] Orthographic and contextual features showed relatively smaller effects, suggesting that they were largely redundant with representations learned by the encoder.

Group Removed	es RMSE	Δ
None (all features)	1.092	—
− Frequency/norms	1.104	+0.012
− Cognate	1.100	+0.008
− Semantic	1.066	−0.026
− Orthographic	1.088	−0.004
− Context	1.046	−0.046

Table 2: Ablation on Spanish (single seed).

5.2 SHAP analysis

We analyzed feature importance using SHAP values computed on the MLP head. For Spanish and German, age of acquisition score and SUBTLEX-US log frequency emerged as the most influential features, followed by CEFR proficiency level. For Mandarin, stroke complexity was the most important feature by a substantial margin (mean $|\text{SHAP}| = 0.078$).

5.3 Cross-L1 generalization

We evaluated cross-L1 generalization by training XGBoost on one L1's features and testing them on others. Models trained on Spanish and German transferred reasonably well to each other, with only modest performance degradation (+0.02–0.14). In contrast, models trained on Mandarin transferred poorly to European L1s, with a substantial increase in RMSE (+0.52).

6 Conclusion

Our findings highlight four main points. First, combining a multilingual encoder with psycholinguistic features yielded consistent improvements over the

[5]One possible explanation is that the transformer already captures orthographic similarity from the joint L1–L2 input, reducing the added value of explicit cognate features for German.

baseline across all L1 groups, with RMSE reductions of approximately 20–26%.

Second, feature contributions varied by L1. While handcrafted features were generally informative, particularly for Spanish and German, their impact differed across languages, suggesting that vocabulary difficulty reflects L1-specific patterns.

Third, cross-L1 transfer results provided additional evidence for this pattern: models trained on Spanish and German tended to generalize well to each other, whereas transfer to and from Mandarin was substantially weaker. From a practical perspective, these findings indicate that effective systems should balance generalizable features with L1-specific adaptations, particularly when extending to typologically distant learner populations.

Fourth, not all features were beneficial when paired with a strong encoder. Ablation results showed that removing contextual and orthographic features could improve performance, indicating that such information is already captured by the pretrained transformer. This underscores the importance of selectively incorporating features that provide complementary signals beyond what the model can learn from text alone.

7 Limitations

This study has several limitations. First, hyperparameter tuning was conducted only for Spanish due to computational constraints; as a result, the German and Mandarin models may not be optimally calibrated. Second, we did not evaluate a standalone XLM-RoBERTa-large baseline without feature augmentation, limiting our ability to isolate the contribution of the feature branch. Third, ablation experiments were conducted with a single random seed, which may introduce variability in the estimates.

In addition, attempts to use `mdeberta-v3-base` as an alternative encoder resulted in unstable training (NaN losses) across all L1 groups, and time constraints prevented further investigation. Finally, for words not covered by external psycholinguistic resources, median imputation was applied, which may degrade performance on rare or out-of-vocabulary items.

References

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of KDD*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

Council of Europe. 2001. Common European framework of reference for languages.

Scott Crossley, Kristopher Kyle, and Thomas Salsbury. 2016. A usage-based investigation of l2 lexical acquisition: The role of input and output. *The Modern Language Journal*, 100(3):702–715.

Ton Dijkstra, Koji Miwa, Bianca Brummelhuis, Maya Sappelli, and Harald Baayen. 2010. How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and Language*, 62(3):284–301.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv:2002.06305*.

Mariano Felice and Lucy Skidmore. 2026. Findings of the bea 2026 shared task on vocabulary difficulty prediction for english learners. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Minkyung Kim, Scott A Crossley, and Kristopher Kyle. 2018. Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, 102(1):120–141.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4):978–990.

Batia Laufer. 1997. What's in a word that makes it hard or easy? In *Vocabulary: Description, Acquisition and Pedagogy*, pages 140–155.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

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

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

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

Terence Odlin. 2003. Cross-linguistic influence. *The handbook of second language acquisition*, pages 436–486.

Gustavo H. Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francesco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Håkan Ringbom. 2007. *Cross-linguistic Similarity in Foreign Language Learning*. Multilingual Matters.

Job Schepens, Ton Dijkstra, and Franc Grootjen. 2012. Distributions of cognates in europe as based on levenshtein distance. *Bilingualism: Language and Cognition*, 15(1):157–166.

Norbert Schmitt, Karen Dunn, Barry O'Sullivan, Laurence Anthony, and Benjamin Kremmel. 2024. *Knowledge-based Vocabulary Lists*. University of Toronto Press, Toronto.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Matthew Shardlow and 1 others. 2024. Bea 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Hua Shu, Xi Chen, Richard C. Anderson, Ningning Wu, and Yan Xuan. 2003. Properties of school chinese: Implications for learning to read. *Child Development*, 74(1):27–47.

Lucy Skidmore, Mariano Felice, and Karen J. Dunn. 2025. Transformer architectures for vocabulary test item difficulty prediction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 160–174, Location TBD. Association for Computational Linguistics.

Hakyung Sung, Mikyung Kim Wolf, Michael Suhan, and Kristopher Kyle. 2025. Lexical richness in young english learners' writing: A focus on opinion and listen-write task types. *Assessing Writing*, 66:100975.

Unicode Consortium. 2024. The unicode standard, version 16.0: Unihan database. `https://www.unicode.org/standard/standard.html`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Timothee Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

David H. Wolpert. 1992. Stacked generalization. *Neural Networks*, 5(2):241–259.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Henrique Paetzold, Lucia Specia, Svenja Ser, Lea A. Deleris, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2018)*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.