UGA Threshold at BEA 2026 Shared Task 1: Predicting Vocabulary Acquisition Difficulty with Hand-Crafted SLA-Based Features

Emma Dalbo
University of Georgia
emma.dalbo@uga.edu

Abstract

This paper describes a feature-based system submitted to the BEA 2026 Shared Task on Vocabulary Difficulty Prediction (closed track). The system models vocabulary difficulty for English learners using linguistically motivated features capturing frequency, cross-linguistic similarity, phonological and orthographic complexity, and semantic properties, supplemented by multilingual embeddings (reduced via PCA). Multiple regression models were evaluated using cross-validation, with final predictions generated from ensemble and single-model configurations per language.

The system achieves competitive performance across all three L1 groups (German, Spanish, and Chinese), outperforming the XLM-RoBERTa baseline in seven of nine runs in terms of RMSE, with the strongest gains observed for Chinese and more modest improvements for Spanish. An ablation study further demonstrates that frequency and cross-linguistic similarity factors contribute most substantially to predictive performance, with effects varying across L1s. These findings highlight the role of interpretable linguistic features in modeling vocabulary difficulty in an L1-aware setting.

1 Introduction

The BEA 2026 Shared Task on Vocabulary Difficulty Prediction asks participants to predict how difficult it is for second language (L2) learners of English to recall a target word in a controlled lexical access task. In each item, learners were presented with a sentence in their native language (L1), a translation equivalent in that L1, and a partial spelling clue for the English target word (consisting of the first letter followed by underscores for hidden letters). The target variable for this task is a GLMM-derived difficulty score computed from large-scale learner response data, with more negative values indicating greater difficulty.

Vocabulary knowledge is a central component of language proficiency, underpinning both comprehension and production. Accurately estimating vocabulary difficulty is therefore critical for applications such as adaptive testing, curriculum design, and personalized learning. Prior work in lexical complexity prediction has primarily focused on general-language audiences (North et al., 2023), with comparatively less emphasis on modeling learner-specific factors such as L1 background. However, cross-linguistic transfer, frequency effects, and psycholinguistic properties of words are known to influence acquisition in systematic ways.

The shared task extends this line of work by framing vocabulary acquisition as an L1-aware regression problem. Separate datasets were provided, representing German (de), Spanish (es), and Chinese (cn) learners respectively. Submissions were evaluated using root mean squared error (RMSE) and Pearson correlation against held-out test data.

The baseline system provided by task organizers fine-tuned XLM-RoBERTa-base on concatenated raw input fields, treating the problem as a sequence regression task without explicit feature engineering. In contrast, this work adopts an approach grounded in theoretically motivated factors from second language acquisition (SLA) and psycholinguistic research, using interpretable hand-crafted inputs in vocabulary difficulty prediction, rather than a complex fine-tuned black box-style model. This system participates in the closed track, which restricts the use of external training data and prohibits large language models. All external resources, libraries, and implementation details are documented in the accompanying supplementary material.

2 System Description

2.1 Task Formulation

Each item in the given dataset includes the following fields: a unique item identifier, the learner's

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 992–996

L1, the English target word and its part of speech, a partial spelling clue, an L1 translation equivalent, an L1 sentence context, and a GLMM-derived difficulty score. The task is framed as supervised regression over this data.

Each language-specific training set contains 6,091 items, with 677 development items and 748 test items. Models are trained separately for each L1.

2.2 Preprocessing

Several preprocessing steps are applied prior to feature extraction. Parenthetical content in both the L1_source_word and L1_context fields is removed using regular expressions, since these represent annotations rather than learner-visible input. When multiple L1 translation equivalents were provided in a comma-separated list, only the first item is used for feature computation.

The L1 context is normalized by removing extra whitespace, and the English target word is lowercased for all lookup-based feature computations. Additional intermediate fields (e.g. cleaned L1 context and processed L1 source forms) are constructed to support feature extraction.

2.3 Feature Engineering

The feature set is designed to prioritize interpretability and is motivated by established findings in SLA and psycholinguistics. These features aim to capture multiple dimensions of vocabulary difficulty, including exposure frequency, cross-linguistic similarity, phonological complexity, and semantic properties. All features are computed from the provided data and publicly available resources. The selected feature groups reflect a balance between theoretical motivation and feasibility under the constraints of the shared task setting.

Frequency-based features. Word frequency is a well-established predictor of lexical acquisition (Ellis, 2002). Frequency scores (Zipf scale) are computed for both the English target word and its L1 translation using the wordfreq library. A feature representing the difference between the two frequencies is also included to capture relative familiarity across the languages.

Cross-linguistic similarity. Cross-linguistic similarity can facilitate vocabulary acquisition through transfer effects, particularly in the case of cognates (Otwinowska and Szewczyk, 2017). This is modeled using multiple similarity measures between the English target word and the L1 translation, including Levenshtein distance, normalized edit distance, and character overlap ratio. These features approximate orthographic proximity between L1 and L2 forms.

Phonological and orthographic complexity. Aspects of complexity, including phonological and orthographic, influence both perception and production difficulty (Rosa and Eskenazi, 2011). Features include syllable counts, grapheme–phoneme relations, and phonetic complexity measures. Additional features were built to capture consonant and vowel cluster counts, as well as maximum cluster lengths both graphemically and phonemically; this provides a proxy for articulatory and spelling complexity.

Morphological features. Morphological transparency facilitates vocabulary learning, particularly through the presence of common affixes (Carlisle, 2000). We captured this phenomenon through binary indicators of the presence of common prefixes and suffixes, as well as an aggregate affix count.

Contextual features. Contextual information is derived from the provided L1 sentence. Features include sentence length (in words and characters) and both the absolute and relative positions of the target word within the sentence. These features serve as rough proxies for contextual importance and positional prominence. We did not incorporate richer contextual representations such as syntactic structure or dependency-based features due to time and complexity constraints. While part-of-speech information for the target word is included via one-hot encoding, broader contextual linguistic features were not explored.

Linguistic metadata. Part-of-speech information for the English target word is one-hot encoded, allowing the model to capture systematic differences in difficulty across part-of-speech categories. L1 identity is also one-hot encoded, though final models are trained separately per language, effectively eliminating the contribution of this feature.

Semantic features. Concreteness has been shown to influence word learnability (Paivio and Begg, 1971; Brysbaert et al., 2014). We use Brysbaert et al.'s concreteness norms to encode this dimension. Words not present in these norms are assigned the dataset mean as a fallback estimate.

Embedding features. Dense multilingual embeddings are generated for both the English target word and the L1 translation using the paraphrase-multilingual-MiniLM-L12-v2 sentence-transformer model. To reduce dimensionality and mitigate overfitting, Principal Component Analysis (PCA) was applied separately to English and L1 embeddings, retaining 50 components each. This dimensionality was selected based on development set experiments as a balance between preserving variance and limiting feature dimensionality, while preventing embedding features from dominating the interpretable feature set. A cosine similarity score between the original embeddings is also included as a feature, representing semantic proximity between the forms.

2.4 Feature Assembly

All feature groups are concatenated into a single feature matrix. Boolean features are converted to integer values. Missing values are filled using column medians computed on the training set (to avoid data leakage), and feature columns are aligned across training, development, and test sets to ensure consistency. The modeling approach intentionally emphasizes relatively simple regression models and interpretable feature design, prioritizing transparency over architectural complexity.

3 Experiments

3.1 Models

Six regression models were evaluated: Ridge regression, SGD regression with ElasticNet regularization, Random Forest, Gradient Boosting, Extra Trees, and a Multilayer Perceptron (MLP), all implemented using scikit-learn.

Each model was wrapped in a pipeline that utilized standard scaling, and hyperparameters were tuned using grid search with 10-fold cross-validation on the training set, optimizing for RMSE. Model selection was performed based on evaluation on the development set.

3.2 Submissions

For all three languages, three runs were submitted:

- **Run 1**: an ensemble formed by averaging the predictions of the four best-performing models on the development set (Gradient Boosting, Extra Trees, Ridge regression, and Multilayer Perceptron), selected based on lowest RMSE.
- **Run 2**: a Gradient Boosting model with fixed hyperparameters.
- **Run 3**: a two-model ensemble combining Gradient Boosting with a secondary model (Extra Trees for Spanish and Chinese; MLP for German).

Final models were retrained on the combined training and development sets before generating predictions for the test set.

4 Results

Table 1 reports the system's performance on the test set for all three closed-track languages, alongside the organizer-provided XLM-RoBERTa baseline.

The system matches or outperforms the XLM-RoBERTa baseline in several conditions: it achieves lower RMSE for all Chinese and German runs and for the ensemble run in Spanish, while Pearson correlation results are more variable across languages. Improvements are most pronounced for Chinese, where all runs outperform the baseline by a substantial margin. For German, all runs achieve a lower RMSE than the baseline. For Spanish, only the ensemble run makes an improvement over the baseline.

Pearson correlation results are more variable across languages. While the system exceeds the baseline's Pearson r for Chinese and one German run, it falls below on Spanish. This suggests that although the system reduces absolute prediction error, the baseline preserves relative ordering of item difficulty more consistently on the Spanish dataset. Performance differences between runs are relatively small, indicating that predictions are relatively stable across model configurations.

5 Analysis and Discussion

The results suggest that a substantial portion of the difficulty signal in this dataset is captured using linguistically interpretable features. Factors such as word frequency, concreteness, grapheme-phoneme correspondence, and cross-linguistic orthographic divergence contribute substantially to predictive performance.

Performance is strongest for Chinese, which may reflect the greater typological distance between Chinese and English. In this case, cross-linguistic similarity features provide clearer signals, and English word-level properties may dominate prediction. For Spanish and German, where cognate rela-

L1	Run	RMSE	Pearson r
Spanish	1	1.236	0.758
	2	1.268	0.740
	3	1.261	0.746
	Baseline	1.257	0.765
German	1	1.176	0.757
	2	1.181	0.753
	3	1.166	0.761
	Baseline	1.258	0.773
Chinese	1	1.031	0.791
	2	1.072	0.770
	3	1.061	0.778
	Baseline	1.140	0.753

Table 1: Performance on the test set across all languages and submissions, including the Baseline for reference. Lower RMSE is better; higher Pearson r is better.

tionships and shared morphological structures are more common, difficulty patterns may be more variable and harder to capture with the current feature set.

To better understand feature contributions, a small-scale ablation study was conducted using the Gradient Boosting model, which was consistently high-performing across languages. Feature groups were removed one at a time, with models retrained on the training set and evaluated on the development set.

Table 2 presents the results of feature-group ablations across all three languages. Overall, frequency-based features were the most critical, with their removal resulting in the largest performance degradation for all languages. This effect was most pronounced for Chinese (ΔRMSE = +0.88), suggesting that frequency plays an especially dominant role for Chinese learners where limited cross-linguistic overlap makes exposure-based factors more central to vocabulary acquisition.

Cross-lingual similarity features also contributed substantially, particularly for German (ΔRMSE = +0.45). The effect was smaller for Spanish and Chinese (ΔRMSE = +0.11 / +0.07).

Embedding-based features provided consistent but secondary improvements across all languages (ΔRMSE $\approx$ +0.18–0.21), indicating that dense representations capture additional signal beyond handcrafted features but do not dominate model performance. In contrast, phonological and orthographic features contributed smaller gains, suggesting a more limited role in this dataset. These results should be interpreted as indicative rather than definitive, given the ablation's limited scope.

These findings are broadly consistent with prior work in SLA and psycholinguistics, while also highlighting the importance of L1-specific effects. However, they should be interpreted within the shared task context, where the objective is to approximate GLMM-derived scores rather than directly model learner acquisition processes.

A brief inspection of high-error items suggests that the model struggles most with low-frequency words and items with limited cross-linguistic similarity, where fewer reliable cues are available. Errors were also more pronounced for Spanish and German items involving partial cognates or morphologically related forms, where orthographic similarity does not consistently correspond to ease of acquisition. In contrast, performance for Chinese appears more strongly driven by frequency-based effects, with fewer ambiguous similarity patterns. These observations indicate that while the feature set captures general trends, it is less effective in cases where multiple competing factors influence difficulty.

Conclusion

This paper presented a feature-based system for L1-aware vocabulary difficulty prediction in the BEA 2026 shared task. The approach combines linguistically motivated features with standard machine learning regression models, achieving competitive performance across all three languages and outperforming the baseline in most conditions in terms of RMSE.

These results suggest that features grounded in established linguistic and psycholinguistic factors, particularly frequency and cross-linguistic similarity, capture a substantial portion of the predictive signal in this dataset. Ablation experiments further

Model Variant	RMSE (de)	RMSE (es)	RMSE (cn)	Avg ΔRMSE
Full model	1.289	1.407	1.233	0.000
– Frequency	1.714	1.789	2.112	+0.56
– Cross-lingual	1.735	1.513	1.299	+0.21
– Embeddings	1.465	1.619	1.419	+0.19
– Phon/Orth	1.323	1.618	1.287	+0.10

Table 2: Feature-group ablation results using the final Gradient Boosting model. Values represent RMSE on the development set (lower is better). ΔRMSE indicates the average increase in error relative to the full model across all languages.

highlight the relative importance of these feature groups and their variation across L1s.

While the modeling approach is intentionally straightforward, the findings contribute to a clearer understanding of how interpretable features relate to vocabulary difficulty in an L1-aware setting. Future work could explore integrating these features with contextualized neural models, improving language-specific representations, and extending the approach to more diverse learner populations.

Ethics Statement

This work uses shared-task data for research purposes. The system predicts aggregate item-level difficulty scores and is not intended for high-stakes decisions about individual learners. While the feature set emphasizes interpretability and theoretical grounding, results may reflect properties of the dataset rather than stable characteristics of learner populations.

Limitations

Several limitations should be noted. Due to computational and time constraints, the scope of model exploration and feature refinement was limited. Although an ablation study was conducted, further analysis could provide deeper insight into feature interactions and redundancy.

Some feature implementations relied on approximations. For example, Chinese syllable count and character length features assumed one syllable and 3 phonemes per character. Additionally, heuristic methods that were used for locating the L1 target word within the context may fail for highly inflected forms.

Context features used were relatively shallow and do not capture the syntactic structure or any deeper semantic relationships. The model also treats each item independently and does not account for individual learner variation past L1 identity.

Finally, the task itself involves predicting GLMM-derived scores, rather than directly modeling inherent word difficulty for human learners. This means the system is approximating an existing statistical model, rather than core learner difficulty patterns.

Acknowledgments

The author thanks Dr. Ryan Ka Yau Lai for his guidance and support, as well as the BEA 2026 organizers and reviewers for their feedback.

References

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.

Joanne F. Carlisle. 2000. Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing*, 12(3):169–190.

Nick C. Ellis. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24(2):143–188.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.

Agnieszka Otwinowska and Jakub M. Szewczyk. 2017. The more similar the better? factors in learning cognates, false cognates and non-cognate words. *International Journal of Bilingual Education and Bilingualism*, 22(8):974–991.

Allan Paivio and Ian Begg. 1971. Imagery and comprehension latencies as a function of sentence concreteness and structure. *Perception & Psychophysics*, 10(6):408–412.

Kevin Dela Rosa and Maxine Eskenazi. 2011. Effect of word complexity on l2 vocabulary learning. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, IUNLPBEA '11, page 76–80, USA. Association for Computational Linguistics.