RETUYT-INCO at BEA 2026 Shared Task 1: Feature-Enriched mDeBERTa for Word Difficulty Prediction

Santiago Robaina and **Luis Chiruzzo** and **Aiala Rosá**
Instituto de Computación, Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay

Abstract

We describe the RETUYT-INCO participation in the BEA 2026 Shared Task on Vocabulary Difficulty Prediction for English Learners, a regression task that predicts GLMM psychometric difficulty scores for English target words given an L1 cue (Spanish, German, and Mandarin). We submitted two systems to the closed track (which restricts participants to the provided shared-task data and standard NLP resources, excluding external corpora and large language models): a feature-engineered XGBoost regressor for all three L1s, and, for Spanish, a 3-seed ensemble of `mdeberta-v3-base` fine-tuned with the same handcrafted features prepended as input text tokens. Our best test result is 1.094 RMSE on Spanish (ensemble), a 13.0% reduction over the XLM-RoBERTa-base closed baseline. We highlight two findings. First, a LaBSE cross-lingual cosine between the L1 source word and the English target word is the largest single-feature addition in our incremental ablation, reducing average development-split (dev) RMSE by 0.091 on top of an already strong string/frequency/POS feature set. Second, feature-only XGBoost, with no neural fine-tuning and no GPU, already beats the XLM-RoBERTa-base closed-track development baseline on average across the three L1s (1.273 vs. 1.287 RMSE).

1 Introduction

Predicting how hard an English word is for a second-language learner is a core problem for intelligent tutoring, adaptive reading, and vocabulary tools. The BEA 2026 Shared Task on Vocabulary Difficulty Prediction for English Learners (Felice and Skidmore, 2026) casts this as a regression problem: given an English target word and an L1 cue (source-language word, context sentence, and a partial orthographic clue), predict a continuous Generalized Linear Mixed Model (GLMM) psychometric difficulty score. The task covers three L1s (Spanish, German, and Mandarin) and two tracks (closed and open). The closed-track baseline is a fine-tuned XLM-RoBERTa-base (Conneau et al., 2019).

The task sits at the intersection of two traditions. On the NLP side, the Complex Word Identification shared tasks (Paetzold and Specia, 2016; Yimam et al., 2018) framed word-level difficulty as classification by non-native readers; SemEval 2021 (Shardlow et al., 2021) replaced this with a continuous Lexical Complexity Prediction target, and BEA 2024 MLSP (Shardlow et al., 2024) extended the setting to multilingual lexical simplification. Handcrafted features very similar to the ones we use here (word length, frequency, morphology, character n-grams) have been standard baselines across this line of work, with fine-tuned neural encoders more recently taking the lead. On the psychometric side, vocabulary testing research long predates these benchmarks (Laufer and Goldstein, 2004), and the Knowledge-based Vocabulary Lists project (Schmitt et al., 2021, 2024) provides the response data that underlies the BEA 2026 targets; the GLMM difficulty scores follow the random-item IRT tradition (De Boeck, 2008; Dunn, 2024). The most direct predecessor is Skidmore et al. (2025), who benchmark transformer architectures for predicting vocabulary-test item difficulty on a KVL-derived corpus; their XLM-RoBERTa setup is adapted as the BEA 2026 closed-track baseline.

We approach the task under a self-imposed lightweight constraint: we target systems that can be trained and served on a single low-end GPU or on CPU, and we restrict ourselves to the closed track. This reflects the compute realities of the Global South, where access to large clusters and API-gated models is limited, and the privacy constraints of on-device educational deployments. We submit two systems for the closed track: (i) an XGBoost (Chen and Guestrin, 2016) regressor over a set of handcrafted string, morphological, POS, fre-

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1113–1118

quency, and cross-lingual semantic-similarity features, for all three L1s, and (ii) for Spanish only, a 3-seed ensemble of `mdeberta-v3-base` (He et al., 2023) fine-tuned for regression with the computed features prepended as input text tokens.

Our research question centres on which handcrafted features carry the most signal for cross-lingual vocabulary difficulty prediction. Two findings emerge. First, a LaBSE (Feng et al., 2022) cross-lingual cosine between the L1 source word and the English target word is the largest single-feature drop we observe in an incremental-addition ablation (Table 1), reducing average development-split RMSE by 0.091 across ES/DE/CN in our tree-based pipeline (§3) and adding a further 0.017 ES development-split RMSE when also prepended as an input token to mDeBERTa (§4). We note that incremental ablations attribute shared variance to whichever feature is introduced first, so this should be read as a lower bound on its marginal contribution rather than a proof of strict dominance over every other single feature; see §3. Second, feature-only XGBoost over our full feature set beats the XLM-RoBERTa-base closed-track development baseline on average (1.273 vs. 1.287), with no neural fine-tuning and no GPU. The official-test picture is mixed: XGBoost beats the test baseline on Mandarin, is within noise on German, and underperforms on Spanish (§5). The mDeBERTa ensemble is reported as our submitted best system, not as an efficiency contribution: it is the same total size as the XLM-RoBERTa baseline per seed, and approximately triples inference cost as an ensemble.

2 Dataset and Task

The task provides, per L1, 6,091 training items and 677 development items, with a held-out test set released at submission time (Felice and Skidmore, 2026). Each item has an English target word and its part of speech, a partial orthographic clue of the target, an L1 translation, and an L1 context sentence. The target is a continuous GLMM psychometric difficulty score, ranging roughly from -6 to $+5$ with lower values indicating harder items. The primary metric is RMSE, with Pearson correlation as a secondary metric. Two tracks are defined: closed (only the provided data and standard NLP resources) and open (external data and LLMs allowed). We participate in the closed track only; the open track was out of scope under our lightweight constraint.

3 Feature Engineering and XGBoost

3.1 Feature set

We extract a set of handcrafted features per item, grouped in this section by type rather than by order of addition. All features are computed on CPU.

Target-word features (English only): word length, one-hot POS over the eight task categories, Zipf word frequency from the `wordfreq` package (Speer, 2022), indicators for common English suffixes (`-tion`, `-sion`, `-ing`, `-ence`, `-ment`, `-able`, `-ness`, `-ly`) and prefixes (`dis-`, `re-`, `un-`, `sub-`, `out-`, `over-`), maximum consonant cluster length, double-letter indicator, syllable count, and an English latinate-root indicator.

Source-word features (L1 only): source-word length, source-word syllable count, multiword-source indicator, and an accent indicator that matches Spanish and German diacritics (zero for CN).

Cross-lingual string similarity (both words): length difference and ratio, normalized Levenshtein distance, character-bigram Jaccard overlap, Jaro–Winkler similarity (Winkler, 1990), shared prefix and suffix ratio, a first-letter match indicator, and a suffix-transform match. The suffix-transform rules are hand-written per L1 and encode the expected cognate shape (e.g. *-tion* $\rightarrow$ *-ción* in ES, 15 rules total; *-tion* $\rightarrow$ *-ierung*, *-ty* $\rightarrow$ *-tät*, ... in DE, 18 rules total; empty for CN).

Cross-lingual semantic similarity: `embed_cosine`, the cosine similarity between LaBSE (Feng et al., 2022) sentence embeddings of the English target word and the L1 source word. This is the single largest RMSE drop in our incremental feature ablation (§3) and is the main narrative thread of this paper.

Context: length of the L1 context sentence in characters.

3.2 Model and protocol

We fit an XGBoost regressor on the training split and report RMSE on the development split. Rows 1–3 of Table 1 use linear regression as a deliberately simple baseline for small, near-monotone feature sets; from row 4 onward the features introduce non-linear effects and interactions that motivate XGBoost (`n_estimators`=300, `max_depth`=5, `learning_rate`=0.05), which is also the model used in our submission. The final

row (row 7) uses L1-tuned XGBoost parameters from an 80-iteration random search with 5-fold CV on the training split; tuning moved dev RMSE by only ∼0.001, so feature engineering dominated.

3.3 Observations

Table 1 walks up our feature set. The largest single-feature *incremental* addition is `embed_cosine`, which drops average dev RMSE by 0.091 (row 5 → row 6). The effect is largest on Mandarin (CN dev RMSE 1.257 → 1.181, −0.076), which is consistent with CN having no meaningful orthographic signal between the L1 and English; LaBSE provides the missing cross-lingual bridge. A caveat on attribution: row 5 bundles three features (Jaro–Winkler, char n-gram overlap, Zipf frequency) and produces a larger combined drop (-0.151) than `embed_cosine` alone; because incremental-addition ablations credit shared variance to whichever feature is introduced first, we cannot rule out that one of those three, introduced last, would match or exceed `embed_cosine`'s marginal effect. What we can say is that `embed_cosine` retains a 0.091 average improvement after a feature set that already captures cognate orthography and frequency—i.e., it is not redundant with them. On ES and DE, `embed_cosine` still gives the largest single drop *in this ordering*, alongside target-word Zipf frequency and cross-lingual Jaro–Winkler similarity, which exploit cognate structure.

By row 7, feature-only XGBoost reaches 1.273 average dev RMSE, below the XLM-RoBERTa-base closed dev baseline (1.287), on CPU and with no neural fine-tuning. A leave-one-out ablation on the 9 ES-targeted v2 features confirmed each contributes positively (individual deltas 0.0004–0.0057 on 5-fold CV); we omit per-feature details for space. A full drop-one ablation of `embed_cosine` from the v2 feature set would tighten this claim and is left for future work (§7).

[1]Tuned parameters for row 7: ES uses `n_est=600`, `depth=3`, `lr=0.02`, `subsample=0.85`, `colsample=0.70`, `alpha=0.5`, `lambda=2.0`; DE and CN use `n_est=400`, `depth=3`, `lr=0.05`, `subsample=0.85`, `colsample=0.85`, `alpha=0.1`, `lambda=2.0`. The row 7 / row 6 delta therefore conflates the v2 feature block with the L1-tuned hyperparameters. The v2 feature block itself is applied to all three L1s, with per-L1 rule sets (15 suffix-transform rules for ES, 18 for DE, empty for CN so the suffix-transform component is inert there; the remaining v2 features apply equally to all three L1s).

4 mDeBERTa with Feature-Enriched Input

For Spanish, we fine-tune `microsoft/mdeberta-v3-base` (He et al., 2023) for regression (`num_labels=1`). The full model is approximately 276M parameters (an 86M transformer backbone plus a ∼190M 250K-token word-piece embedding matrix), essentially the same total size as the XLM-RoBERTa-base closed baseline (∼270M).

Input format. We prepend five feature tokens to the encoder input, then concatenate the raw L1 and English fields with `[SEP]` markers:

```
wlen=N | nedit=N | pos=X | clue=N |
esim=N | L1_word [SEP] L1_context [SEP]
en_clue [SEP] en_word
```

The five prepended features (English word length, normalised edit distance, POS, clue length, and `esim`) are a subset of the XGBoost feature set rendered as plain text. `esim` is the LaBSE cosine of §3 (`embed_cosine`), now available to the encoder as a numeric input token.

Training. GLMM scores are normalised to zero mean and unit variance at training time and denormalised before RMSE computation. We train with learning rate 2×10^{-5}, a cosine schedule with 10% warmup, up to 10 epochs with early stopping (patience 3), batch size 32, weight decay 0.01, fp16, and max sequence length 256. Training was performed on a single Google Colab T4 GPU.

Ensemble. We train three seeds ($\{10, 42, 123\}$) and average their predictions. The ensemble triples inference compute and memory: three ∼276M-parameter models are loaded in parallel, for approximately 828M parameters and 3× forward passes per prediction. Each individual seed already beats the XLM-RoBERTa-base baseline on ES dev (Table 2) by 0.179–0.220 RMSE; the ensemble is a further refinement of 0.034 RMSE over the best individual seed (seed 42, 1.137) and 0.059 RMSE over the mean of the three seeds, so it is a refinement on top of the per-seed gain rather than the source of the gap to the baseline.

Why Spanish only. We did not extend the mDeBERTa ensemble to all three L1. DE and CN ensemble were skipped for time reasons and our own limitations of knowledge of the languages. Our DE and CN submission is therefore XGBoost only.

Test-time retraining. The submitted ES ensemble is retrained on `train+dev` combined before predicting the test set. The test RMSE (1.094) is

Feature set	Model	ES	DE	CN	Avg	Δ vs. prev
1. `word_len`	LR	1.778	1.707	1.509	1.665	—
2. + edit distance	LR	1.705	1.596	1.487	1.596	−0.069
3. + POS	LR	1.663	1.536	1.493	1.564	−0.032
4. + morphology + shared prefix + context	XGBoost	1.613	1.524	1.459	1.532	−0.032
5. + Jaro–Winkler + n-gram + Zipf freq.	XGBoost	1.464	1.421	1.257	1.381	−0.151
6. **+ `embed_cosine`** (LaBSE)	XGBoost	**1.350**	**1.339**	**1.181**	**1.290**	**−0.091**
7. + L1-specific (v2)	XGBoost[1]	1.327	1.334	1.158	1.273	−0.017
XLM-RoBERTa-base (closed) baseline	XLM-R (∼270M)	1.357	1.328	1.175	1.287	—

Table 1: Feature progression on dev (RMSE, 3 dp). Rows 1–3 add the target-word length, a first cross-lingual string feature (edit distance), and target POS. Row 4 adds target-word morphology (suffix/prefix indicators, consonant cluster, double-letter) together with the shared-prefix ratio and context length. Row 5 adds further cross-lingual string similarity (Jaro–Winkler, character-bigram overlap, first-letter match) and target-word Zipf frequency. Row 6 (bold) adds the cross-lingual semantic similarity (`embed_cosine`). Row 7 adds the L1-specific v2 block and switches to L1-tuned XGBoost hyperparameters. The last row of the table is the task baseline for reference.

therefore not directly comparable to the dev RMSE of the same configuration (1.103).

5 Results and Analysis

On Dev-set. Feature-only XGBoost beats the XLM-RoBERTa-base closed dev baseline on average and on ES and CN individually; on DE it is within noise of the baseline (1.334 vs. 1.328, +0.006). See Table 2. The pipeline uses CPU-fit trees and no neural fine-tuning. The submitted ES ensemble reaches 1.103 dev RMSE, a further 0.224 RMSE below our XGBoost row. This gap is not an efficiency result, since per-seed mDeBERTa has essentially the same parameter count as XLM-RoBERTa-base; it is an effect of the feature-enriched input format, target scaling, and 3-seed averaging. Each individual seed already beats the XLM-RoBERTa dev baseline on ES.

On Test-set. On the official closed-track test leaderboard (Table 3), our ES ensemble reports 1.094 RMSE / 0.843 Pearson, placing us 9th of 21 participating teams on Spanish (best-per-team collapse), and reducing the test-baseline RMSE by 13.0%. XGBoost on Mandarin also beats the test baseline (1.106 vs. 1.140, −3.0%), placing us 18th of 21 teams. German XGBoost is within noise of the test baseline (1.260 vs. 1.258, +0.2%, rank 18/20). ES XGBoost underperforms the test baseline (1.323 vs. 1.257, +5.3%), despite beating the dev baseline (1.327 vs. 1.357). This dev/test inversion on ES is the clearest sign of distribution shift between the two splits for our feature-based system.

6 Conclusions

Among the handcrafted features we tested, the LaBSE cross-lingual cosine between the L1 source word and the English target word carries the most task-relevant signal: adding it reduces average dev RMSE by 0.091 across ES/DE/CN in our XGBoost pipeline, and adds a further 0.017 ES dev RMSE when also included as a prepended input token to mDeBERTa. Feature-only XGBoost, with no GPU, already beats the XLM-RoBERTa-base closed dev baseline on average (1.273 vs. 1.287). Our submitted ES system, a 3-seed mDeBERTa ensemble, places 9/21 on the closed-track ES test leaderboard at 1.094 RMSE.

7 Limitations

The neural ensemble was run for ES only; DE was skipped for compute reasons and CN was skipped by design (Latin-script feature assumptions). The submitted ES ensemble loads ∼828M parameters and runs three forward passes per prediction: it is not an efficient system relative to the baseline, and we report it as our best test result rather than as an efficiency contribution. We prioritise breadth over depth and did not run per-L1 mDeBERTa hyperparameter searches. The ES XGBoost shows a clear dev/test inversion that we do not fully explain. We did not explore open-track resources. Finally, our "largest single-feature incremental drop" claim for `embed_cosine` rests on incremental-addition evidence and is therefore order-dependent: the three features added in row 5 of Table 1 (Jaro–Winkler, char n-gram, Zipf) share credit for a combined

[2]The submitted XGBoost boosters were not persisted; their serialised size on disk was not measured before submission. The model is a gradient-boosted tree ensemble (up to 600 trees of depth 3) trained on ≤23 numeric features.

System	Params (total)	ES	DE	CN
XLM-RoBERTa-base (closed baseline)	∼270M	1.357	1.328	1.175
`full_xgb_v2_embed` (our XGBoost, submitted)	tree ensemble (CPU)[2]	1.327	1.334	1.158
mDeBERTa (seed 10)	∼276M	1.178	—	—
mDeBERTa (seed 42)	∼276M	1.137	—	—
mDeBERTa (seed 123)	∼276M	1.172	—	—
mDeBERTa 3-seed ensemble (submitted, ES only)	**∼828M at inference**	**1.103**	—	—

Table 2: Development-split RMSE (3 dp). Per-seed Pearson on ES is 0.789 / 0.821 / 0.826; the ensemble reaches 0.827. The mDeBERTa backbone (86M) and the XLM-RoBERTa backbone (∼85M) are essentially the same size; total parameters including the 250K × 768 word-piece embedding matrix are ∼276M and ∼270M respectively.

L1	System	RMSE	Pearson	Team rank	Δ RMSE vs. baseline
ES	mDeBERTa 3-seed ensemble (submitted)	1.094	0.843	9 / 21	−0.163 (−13.0%)
ES	XGBoost (secondary submission)	1.323	0.713	—	+0.066 (+5.3%)
DE	XGBoost (submitted)	1.260	0.713	18 / 20	+0.002 (+0.2%)
CN	XGBoost (submitted)	1.106	0.754	18 / 21	−0.034 (−3.0%)
ES	XLM-RoBERTa-base (closed test baseline)	1.257	0.765	—	—
DE	XLM-RoBERTa-base (closed test baseline)	1.258	0.773	—	—
CN	XLM-RoBERTa-base (closed test baseline)	1.140	0.753	—	—

Table 3: Official test-set results, closed track. A negative Δ RMSE indicates our system is lower than (better than) the baseline. Team rank collapses each team to its best submission per L1; dashes in that column indicate a row shown for comparison only and not counted separately (our ES team rank is determined by the ensemble). The ES XGBoost row is reported to contextualise the dev/test behaviour of the feature-only system.

-0.151 drop, and a drop-one ablation from the full v2 set would be needed to establish strict per-feature dominance. We leave that ablation to future work.

Acknowledgments

This paper was funded by *Agencia Nacional de Investigación e Innovación* (ANII, Uruguay), Projects $FSED\_2\_2023\_1\_179355$ and $FMV\_1\_2023\_1\_176581$.

References

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Paul De Boeck. 2008. Random item IRT models. *Psychometrika*, 73(4):533–559.

Karen J. Dunn. 2024. Random-item Rasch models and explanatory extensions: A worked example using L2 vocabulary test item responses. *Research Methods in Applied Linguistics*, 3(3):100143.

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*. To appear.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Batia Laufer and Zahava Goldstein. 2004. Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, 54(3):399–436.

Gustavo 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.

Norbert Schmitt, Karen Dunn, Barry O'Sullivan, Laurence Anthony, and Benjamin Kremmel. 2021. In-

troducing knowledge-based vocabulary lists (KVL). *TESOL Journal*, 12(4).

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

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024. The 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)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

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.

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

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

William E. Winkler. 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.