Token Titans at BEA 2026 Shared Task 1: Multilingual Lexical Complexity Prediction via Fine-Tuned XLM-RoBERTa with Ensemble Decoding

Anubhab Parashar and Sandeep Mathias
Presidency School of Computer Science and Engineering
Presidency University, Bangalore
{anubhab.20231COM0043, sandeepalbert}@presidencyuniversity.in

Abstract

In this paper, we describe our submission to the BEA 2026 Shared Task on Multilingual Lexical Complexity Prediction. Our system fine-tunes XLM-RoBERTa Large separately for Spanish, German, and Chinese, feeding each instance as a concatenation of the source word, its context in a sentence, an English clue, and the English target word. Training uses Z-score label normalization and two independent runs that differ in learning rate, random seed, etc. On the official test set, the system scores RMSE = 1.170 and Pearson Correlation = 0.812.

1 Introduction

Estimating how hard a word is to read depends on who is doing the reading. A word which is easy for a native speaker to understand may be quite complex for a language learner, and the same word can sit at opposite ends of the difficulty scale depending on its context. Lexical complexity prediction (LCP) formalizes this as a regression task: given a word in sentence context, predict a score reflecting how difficult it is for a reader.

Most LCP work has concentrated on English (Shardlow et al., 2021; North et al., 2022). The BEA 2026 Shared Task extends this to learners whose first language (L1) is Spanish, German, or Chinese – three languages that differ substantially in morphology, script, and how well-represented they are in multilingual pretrained models. Chinese is the most challenging of the above languages–no whitespace segmentation, character-level ambiguity, and thinner coverage in most encoders than either of the Latin-script languages.

The BEA 2026 Shared Task on Vocabulary Difficulty Prediction for English Learners requires participants to predict the complexity of a word for learners of English from different L1 backgrounds - namely Spanish, German and Chinese. Each instance in the task also includes an English clue word and an English target alongside the L1 word and context. A model with strong English representations can potentially use the clue and target as anchors even when the source language is less familiar. The dataset in the shared task provides us with the source word, context, clue and target, which can help the classifier in predicting the complexity.

We fine-tune XLM-RoBERTa Large (Conneau et al., 2020) separately for each language, concatenating all four fields into a single flat input string. To squeeze a bit more out of the model without additional data, we train two runs per language under slightly different hyperparameters and improve model performance via weighted averaging of predictions. The final system scores RMSE as 1.170 and Pearson $r = 0.812$ on the test set.

2 Related Work

Lexical complexity prediction. Early readability work used word frequency and syllable counts as proxies for lexical difficulty (Dale and Chall, 1948). The SemEval-2021 LCP shared task (Shardlow et al., 2021) shifted the benchmark toward fine-grained regression on words in context, and transformer-based systems pulled well ahead of feature-engineered pipelines - less because of architectural creativity than because model scale simply helps (North et al., 2022; Zaharia et al., 2022).

Multilingual LCP has received less attention. Most multilingual efforts swap in a multilingual encoder and, sometimes, language-specific lexical resources such as frequency lists or morphological analyzers (Shardlow et al., 2022). An earlier shared task on multilingual lexical simplification was organized during BEA 2024 (Shardlow et al., 2024)

Multilingual pretrained encoders. XLM-RoBERTa (Conneau et al., 2020) covers 100 languages and has become the default backbone

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1119–1123

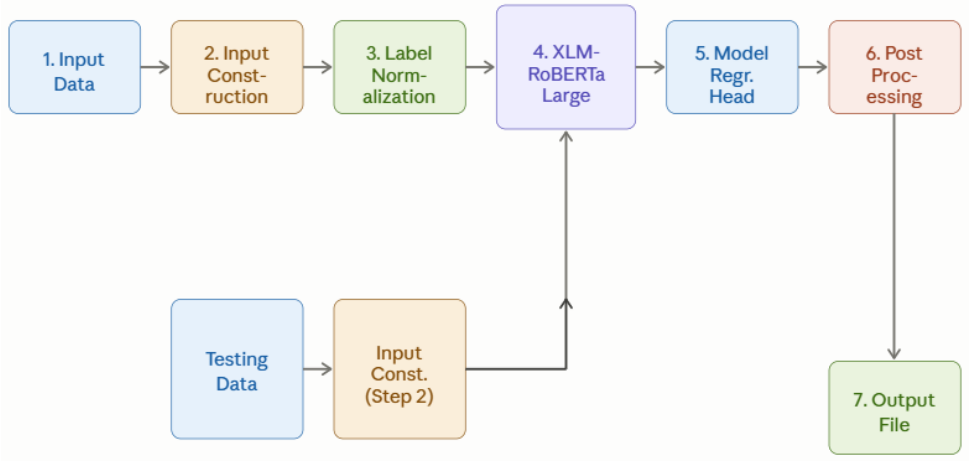

Figure 1: Overview of the proposed pipeline for multilingual lexical complexity prediction.

for cross-lingual classification and regression. In this Shared Task, the standard setup - linear head on `[CLS]`, full fine-tuning - works well enough that there is little reason to deviate from it.

Ensembling. Averaging predictions across runs with different seeds or hyperparameters reduces variance without requiring additional data or a larger model (Lakshminarayanan et al., 2017; Liu et al., 2019). We train two runs per language and combine them with a fixed weighted average (0.6/0.4), with weights chosen on the validation set.

3 System Description

Figure 1 shows our system architecture. For training, we do input construction and label normalization which we feed as input to XLM-RoBERTa to learn our model. At testing time, we also construct the input representations as input to XLM-RoBERTa and the model. The output from the model is sent to a regression head, followed by post-processing to generate our output labels.

3.1 Input Construction

Each instance has four fields: the L1 source word, its sentence context, an English clue word, and an English target word. We join them with `</s>` separators, which XLM-RoBERTa treats as segment boundaries. All four fields are concatenated as:

```
L1_word </s> L1_context </s> en_clue </s>
en_target
```

For Chinese, consecutive whitespace is collapsed with a regex pass before tokenization; without it, the tokenizer splits character-segmented text erratically.

3.2 Model

We fine-tune XLM-RoBERTa Large (Conneau et al., 2020) with a linear regression head on the `[CLS]` token. The full network is updated end-to-end. We train separate models for each language rather than a single joint model; the three languages differ enough in script and morphology that joint training hurt validation RMSE in preliminary experiments.

GLMM scores are z-score normalized before training:

$$\hat{y} = \frac{y - \mu}{\sigma} \tag{1}$$

where μ and σ are computed over the training split for each language. Predictions are denormalized before submission and clipped to $[0.05, 5.0]$; the lower clip handles a small number of predictions that collapsed near zero during early stopping.

3.3 Training

All models use AdamW with fp16 mixed precision, gradient checkpointing, 500 warmup steps, and gradient norm clipped at 1.0. Batch size is 4 with 2 gradient accumulation steps (effective batch size 8), trained for 5 epochs. The checkpoint with the lowest validation RMSE is kept.

3.4 Ensembling

We train two runs per language with different hyperparameters (Table 1). Spanish Run 1 produced the strongest single result in the submission, and its settings were the starting point for German and Chinese - we adjusted learning rate and seed between runs to test whether that configuration would hold. It did not, which is discussed in Section 4.

For German only, we combined Run 1 and Run 2 via weighted averaging:

$$\hat{y}_{\text{ens}} = 0.6 \cdot \hat{y}_1 + 0.4 \cdot \hat{y}_2 \quad (2)$$

The weights were set on the validation set across all languages and not tuned further on a per-language basis.

Lang.	Setting	Run 1	Run 2
es / de	Learning rate	1e-5	1.1e-5
	Scheduler	linear	cosine
	Seed	42	1337
	Max length	384	320
cn	Learning rate	1.2e-5	1.1e-5
	Scheduler	cosine	cosine
	Seed	42	1337
	Max length	320	320

Table 1: Per-language hyperparameters. The following hyperparameters were shared across all runs: AdamW, fp16, batch size 4, 2 gradient accumulation steps, 500 warmup steps, 5 epochs, max_grad_norm 1.0.

3.5 Experimental Setup

We evaluate on the official BEA 2026 test set using RMSE (primary) and Pearson r (secondary). Training and validation splits follow the shared task data release. No additional training data or external resources are used. We compare against the shared task baseline, which fine-tunes `xlm-roberta-base` with batch size 32, learning rate 3×10^{-5}, and dropout 0.1 on the same four input fields; baseline results are reported on the test set.

4 Results and Analysis

Spanish Run 1 achieved the strongest result in the submission (RMSE = 1.170, Pearson Correlation = 0.812). Spanish Run 2 collapsed entirely—Pearson of 0.043 indicates the model's predictions are near-random, likely due to training instability under the cosine schedule with seed 1337 on the Spanish data distribution. This is consistent with the instability findings discussed in Section 4.1; the same seed and scheduler combination that worked for German produced a degenerate run for Spanish. Run 2 was excluded from the Spanish ensemble for this reason.

For German and Chinese, both runs completed normally. Run 1 is the weakest of the two runs for German on Pearson correlation, and for Chinese the cosine-scheduled Run 2 achieves notably higher Pearson (r = 0.777 vs. 0.718) despite slightly worse RMSE. Ensemble predictions were generated for German only, where combining the two runs matched Run 2 exactly (RMSE 1.569, r = 0.772), suggesting the linear-schedule run added no useful signal for that language. The overall best test result is RMSE 1.170 and Pearson r = 0.812 as reported in Table 2.

4.1 Effect of Model Scale

`xlm-roberta-large` has 560M parameters against 270M for the base variant (Conneau et al., 2020). On large fine-tuning sets, the larger model reliably wins. On smaller datasets the picture is less clean–more parameters means more capacity to overfit, and the effect is sharper when the learning rate is low and the batch size is small.

Both conditions apply here. Our effective batch size is 8 (batch 4, gradient accumulation 2) against the baseline's 32, and our learning rates range from 1×10^{-5} to 1.2×10^{-5} against the baseline's 3×10^{-5}. A larger batch size averages out more gradient noise per step, which acts as implicit regularisation; a higher learning rate moves the model away from sharp minima faster. The baseline has both. We have neither—and on German and Chinese, where the training splits are smaller than Spanish, the baseline's implicit regularisation appears to matter more than the capacity advantage of the larger model.

This is consistent with findings in low-resource transfer learning more broadly (Mosbach et al., 2021). Fine-tuning large pretrained models on small datasets is unstable: performance varies significantly across seeds, the best checkpoint is sensitive to when training stops, and the gap between a good run and a bad one is wider than it is for smaller models. Our Chinese results illustrate this directly—Run 1 and Run 2 differ only in learning rate (1.2×10^{-5} vs. 1.1×10^{-5}) and seed, yet their Pearson correlations are 0.718 and 0.777 respectively, a gap of 0.059 on what should be a small hyperparameter change.

One correction that would likely help is adding explicit regularisation to the regression head—dropout before the linear layer, or L2 weight decay on the head parameters specifically. The baseline uses dropout 0.1 throughout. Our runs use no explicit dropout on the head, relying entirely on the pretrained weights and early stopping to prevent overfit. For a 560M parameter model on a few thousand training examples, that is not enough.

Language	System	RMSE ↓	Pearson r ↑
Spanish	Baseline (xlm-roberta-base)	1.257	0.765
	Run 1 (linear, seed 42)	**1.170**	**0.812**
	Run 2 (cosine, seed 1337)	1.885	0.043
German	Baseline (xlm-roberta-base)	**1.258**	**0.773**
	Run 1 (linear, seed 42)	1.564	0.760
	Run 2 (cosine, seed 1337)	1.569	0.772
	Ensemble (0.6/0.4)	1.569	0.772
Chinese	Baseline (xlm-roberta-base)	**1.140**	0.753
	Run 1 (cosine, seed 42)	1.382	0.718
	Run 2 (cosine, seed 1337)	1.436	**0.777**

Table 2: Per-language results compared against the shared task baseline (`xlm-roberta-base`). All figures are on the test set. Bold marks the best result per language per metric.

4.2 Discussion

Table 2 compares our runs against the shared task baseline, which uses `xlm-roberta-base` with a batch size of 32 and learning rate of 3×10^{-5}. The results split clearly by language.

For Spanish, both metrics improve over the baseline—RMSE drops from 1.257 to 1.170 and Pearson rises from 0.765 to 0.812. Upgrading from `xlm-roberta-base` to `xlm-roberta-large` and halving the learning rate proved effective for Spanish, where the training data is likely well represented in the larger model's pretraining corpus.

The baseline RMSE of 1.258 for German is substantially better than our best result of 1.564, and for Chinese, the baseline (1.140) outperforms both our runs on RMSE. On Pearson correlation, the baseline leads in German as well (0.773 vs. 0.772), while for Chinese our Run 2 reaches r = 0.777 against the baseline's 0.753. The primary metric goes to the baseline on both languages.

The inconsistency across languages also points to a deeper issue: the hyperparameters that work for Spanish do not generalise. Spanish benefits from dense coverage in XLM-RoBERTa's pretraining data and relatively transparent morphology; German and Chinese present harder targets for the same configuration.

5 Conclusion and Future Work

The system is straightforward: one pretrained encoder, per-language fine-tuning, a flat input string, two runs averaged together. The system makes no architectural contributions. What the results show is that this setup is sufficient for Spanish—RMSE = 1.170 and Pearson Correlation = 0.812 without external resources, language-specific features, or architectural modifications—but does not generalise cleanly across languages.

The English-side signal warrants further investigation. The clue and target words may provide useful cross-lingual anchors, particularly for Chinese where the encoder's L1 coverage is weakest, and future work should include a controlled ablation to isolate their contribution.

The clearest lesson from the results is that a single hyperparameter configuration cannot serve three typologically different languages equally well. Spanish, German, and Chinese respond differently to learning rate, batch size, and scheduler choice. Future work should tune these per language from the start, add explicit regularisation to the regression head, and retain per-run scores across all languages to enable proper ablation.

In the future, we also plan to adopt the use of LLMs, such as GPT, Claude, etc. for improving the performance of Lexical Complexity Prediction. While the usage of LLMs has been shown to perform quite well in tasks such as generating simplifications across multiple languages ((Aumiller and Gertz, 2022; Dutilleul et al., 2024)), to the best of our knowledge, not much work has gone into measuring lexical complexity across multiple languages.

Limitations

Intermediate per-run predictions were not saved for Spanish and Chinese, preventing us from constructing and evaluating ensemble variants for those languages post-hoc. The ablations mentioned in Section 3 were run on the validation set only and were not systematic; we adjusted one variable at a time without a full factorial search. The ensemble weights (0.6/0.4) were set by inspection across all languages rather than optimized per language—the German results suggest this was suboptimal.

Finally, the system uses no morphological analysis, frequency information, or psycholinguistic norms–features that have historically been strong predictors of lexical complexity and that might close some of the remaining gap.

References

Dennis Aumiller and Michael Gertz. 2022. UniHD at TSAR-2022 shared task: Is compute all we need for lexical simplification? In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 251–258, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28.

Benjamin Dutilleul, Mathis Debaillon, and Sandeep Mathias. 2024. ISEP_Presidency_University at MLSP 2024 shared task: Using GPT-3.5 to generate substitutes for lexical simplification. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 605–609, Mexico City, Mexico. Association for Computational Linguistics.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6402–6413.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496. Association for Computational Linguistics.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.

Sian North, Fernando Alva-Manchego, and Matthew Shardlow. 2022. Predicting multilingual sentence readability with transformer models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3054–3065. International Committee on Computational Linguistics.

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, and 3 others. 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 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. Association for Computational Linguistics.

Matthew Shardlow, Marcos Zampieri, and Richard Evans. 2022. Predicting the complexity of words in context with lexical resources and transformer models. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 91–100. Association for Computational Linguistics.

George-Eduard Zaharia, Radu Ion, and Dan Tufis. 2022. Exploring transformer models for lexical complexity prediction. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 11–20. Association for Computational Linguistics.