Transformer Architectures for Vocabulary Test Item Difficulty Prediction

Lucy Skidmore, Mariano Felice, Karen J. Dunn
English Language Research, British Council, UK
name.surname@britishcouncil.org

Abstract

Establishing the difficulty of test items is an essential part of the language assessment development process. However, traditional item calibration methods are often time-consuming and difficult to scale. To address this, recent research has explored natural language processing (NLP) approaches for automatically predicting item difficulty from text. This paper investigates the use of transformer models to predict the difficulty of second language (L2) English vocabulary test items that have multilingual prompts. We introduce an extended version of the British Council's Knowledge-based Vocabulary Lists (KVL) dataset, containing 6,768 English words paired with difficulty scores and question prompts written in Spanish, German, and Mandarin Chinese. Using this new dataset for fine-tuning, we explore various transformer-based architectures. Our findings show that a multilingual model jointly trained on all L1 subsets of the KVL achieve the best results, with analysis suggesting that the model is able to learn global patterns of cross-linguistic influence on target word difficulty. This study establishes a foundation for NLP-based item difficulty estimation using the KVL dataset, providing actionable insights for developing multilingual test items.

1 Introduction

Calibrating the difficulty of test items is a core aspect of language assessment design, ensuring that tests are fair, consistent, and aligned with learner proficiency. Traditionally, this calibration relies on pre-testing large item samples or expert judgment, which are expensive and time-consuming. Consequently, there is an increasing interest in automating item calibration using machine learning methods (Yancey et al., 2024; Yaneva et al., 2024), which offer greater scalability, efficiency, and consistency, and can be more easily integrated into item development pipelines.

While transformer-based encoder models such as BERT (Devlin et al., 2019) have been successfully applied to question difficulty estimation from text (QDET) in domains related to content knowledge assessment, approaches in language assessment — where difficulty is more closely tied to the linguistic properties of the item — still largely rely on handcrafted features (AlKhuzaey et al., 2024). This is particularly true in QDET for L2 English vocabulary items, which commonly rely on small datasets that are not suitable for fine-tuning transformers (Benedetto et al., 2023).

To address this gap, our paper introduces a new multilingual resource for vocabulary QDET: an extended version of the British Council's Knowledge-based Vocabulary Lists (KVL), containing 6,768 English vocabulary items paired with difficulty scores and prompts in Spanish, German, and Mandarin Chinese. We use the KVL to fine-tune various transformer-based architectures for vocabulary test item difficulty prediction, leveraging its unique structure to provide insights on how best to model vocabulary difficulty in multilingual settings. Our exploratory work serves as a benchmark for future development of generalisable, L1-agnostic models for explainable item calibration.

Our paper begins with an overview of how the KVL has been extended and adapted for NLP-based applications. This is followed by a summary of the latest research in two domains that this study intersects: question difficulty estimation from text and lexical complexity prediction. Next, we outline the aims of the study, providing the motivation and context for the experimental design. The transformer-based architectures that we investigated are outlined, including the procedures followed for model selection and fine-tuning. We present findings from model performance evaluations, an ablation study and an error analysis. Finally, the paper ends with a discussion of the results and outlines directions for future work.

2 The Knowledge-based Vocabulary Lists

The Knowledge-based Vocabulary Lists (KVL) (Schmitt et al., 2021, 2024) were the outcome of a collaborative research project between the British Council and researchers from the University of Nottingham, University of Innsbruck and Waseda University. Productive English language word knowledge was assessed using prompts designed to test form-based recall of individual lemmas in a translation format (cf. Laufer and Goldstein, 2004). Items comprising an L1 translation of the English target word, plus contextualising sentences were developed separately in three L1s (Mandarin Chinese, German, Spanish) to create a bank of 7,679 items for each language. Participants were required to input the remainder of the word in English, as per this example from the Spanish language version[1]:

> casa Vivo en una casa grande que tiene tres dormitorios.
> h _ _ _ _

During a period between late-2018 to mid-2020, 3.3 million responses were collected from over 100,000 respondents via crowdsourcing. An online platform, promoted across the British Council's social media channels, presented participants with blocks of ten random items stratified by target word frequency. Feedback was given after each block, and participants were encouraged to complete more items to "beat their best", as an example of game-based data collection (Kim et al., 2024).

Difficulty estimates were derived separately for each L1 subset of the data, using random-item-random-person (RPRI) Rasch models (De Boeck, 2008) built within a generalised linear mixed model (GLMM) framework (Dunn, 2024). Original KVL project outputs used these estimates to create a rank-order list of the top 5,000 words for each L1.

For this research, we use the existing 5000 items in the KVL and publicly release an additional 1,768 English vocabulary test items for each L1. This extended dataset contains 20,304 items in total (6,768 per L1) and is divided into 80% train (16,242 items), 10% development (2,031 items) and 10% test (2,031 items) sets[2].

[1]The German and Chinese versions had similar, yet distinct prompts, for example in German: "Haus Ich wohne in einem Haus mit Garten." And in Chinese: "房子 我买了一座房子。"

[2]https://www.britishcouncil.org/data-science-and-insights/resources

3 Related Work

3.1 Question Difficulty Estimation from Text

Question difficulty estimation from text (QDET) concerns the prediction of test item difficulty based solely on its textual features. There is growing interest in using QDET for high stakes assessment calibration, given its efficiency and scalability compared to traditional methods (AlKhuzaey et al., 2024). The majority of work in this area explores supervised approaches to QDET, with transformer-based encoder models achieving the best results in recent years (Gombert et al., 2024; Yaneva et al., 2024). There is also a growing interest in unsupervised approaches to the task, using generative models as 'test-takers', extracting their uncertainty as a proxy for human difficulty (Loginova et al., 2021; Uto et al., 2024; Zotos et al., 2025).

Research related to vocabulary-based QDET, however, is relatively limited. Most prior approaches to this task use hand-crafted linguistic features (such as word frequency and word length) as inputs to predictive models (Suyong and Hua, 2018; Settles et al., 2020), with other approaches incorporating embeddings such as word2vec (Ehara, 2018) and GloVe (Susanti et al., 2020). Beyond word-based features, contextual factors such as the similarity between correct answers and distractors in multiple choice vocabulary tests (Susanti et al., 2017, 2020) as well as semantic descriptors from dictionary entries of target words (Nakanishi et al., 2012), have also had limited exploration.

3.2 Lexical Complexity Prediction

Lexical complexity prediction (LCP) is a subfield of complex word identification (CWI), which concerns the automatic detection of complex words from text, primarily for the purpose of text simplification. LCP extends the binary classification used for CWI to form a regression problem, with the goal of predicting a continuous 'complexity' value for a given word. These values are domain-specific, and can range from crowd-sourced perceived complexity ratings to morphosyntactically derived features (North et al., 2023). Different to vocabulary QDET, the input text used for LCP typically involves predicting the complexity of a word in context. Given the format of the KVL dataset, the task of LCP aligns more closely with our investigation than much of the previous work in vocabulary QDET.

The most successful approaches to LCP to date make use of transformer-based architectures (Bani Yaseen et al., 2021; Kelious et al., 2024a). Particularly relevant to this work, however, are investigations into multilingual applications of LCP. Sheang (2019) showed that a multilingual CNN model trained jointly on word embeddings and linguistic features of Spanish, German and English datasets led to improved performance of prior models for Spanish and German. Similarly, Finnimore et al. (2019) found that jointly training models with languages from the same family improved cross-lingual CWI. Zaharia et al. (2020) experimented with multilingual transformers for cross-lingual CWI, showing that XLM-RoBERTa performs best for unseen German or French target words. More recently, LLMs have been explored for unsupervised multilingual LCP, however these approaches did not outperform supervised transformer-based equivalents (Kelious et al., 2024b).

4 Research aims

As described above, the KVL dataset is unique in that it contains multilingual test items (comprising L1 source word, L1 context and EN clue) for the same set of English target words across three L1s. To explore this multi-faceted structure, we defined four transformer-based models for experimentation: (1) individual monolingual models for each test item component; (2) ensembles combining these component-specific models; (3) multilingual models fine-tuned on the full test item text separately for each L1; and (4) a single multilingual model trained on the full test item text across all L1s.

Comparing the performance of these models allowed for multiple avenues of investigation: the influence of test item components on model predictions, the suitability of monolingual versus multilingual models and training data, as well as the effectiveness of different architectures for capturing cross-component and cross-lingual interactions within items. In addition to overall model performance, we were also interested in whether model error revealed potential biases—such as systematic under- or overestimation for particular items. These areas of interest were distilled into three primary research questions for the study:

- How accurately can different transformer-based model architectures predict vocabulary item difficulty for the KVL dataset?
- How do the individual components of the test item contribute to the models' predictions?
- Is there any systematic bias contributing to errors in the best-performing model?

5 Modelling setup

Item difficulty prediction was modelled as a regression task. The target values for prediction were transformations of the GLMM item-level conditional modes. These were inversed to reflect item difficulty (as opposed to 'item easiness' in the original study) and scaled to values between zero and one. Models were fine-tuned with mean squared error (MSE) as the loss function. As the KVL were originally designed for ranking vocabulary difficulty, Spearman's rank correlation coefficient (RHO) was used as the main model evaluation metric. The root mean squared error (RMSE) metric was also calculated to evaluate model fit. Where relevant, statistical significance tests of the models were carried out via bootstrap, using 10,000 iterations and Bias-Corrected and Accelerated (BCa) intervals (Efron, 1987).

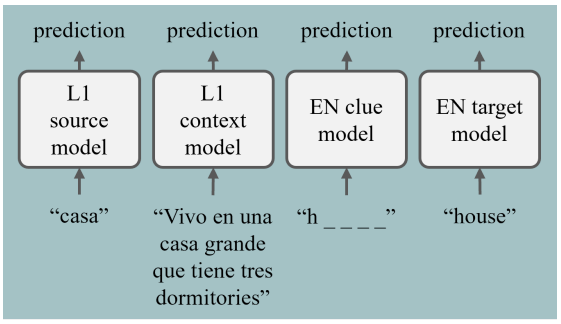
For the multilingual models, the structure of the input text begins with the question content (L1 source word, L1 context, EN clue), in the same order as it is presented in the vocabulary test items, followed by the target answer (EN target word). Each part of the input text was delineated with the models' pre-defined separation token, as shown in the example input text below:

```
casa [SEP] Vivo en una casa grande que
tiene tres dormitorios. [SEP] h____
[SEP] house
```

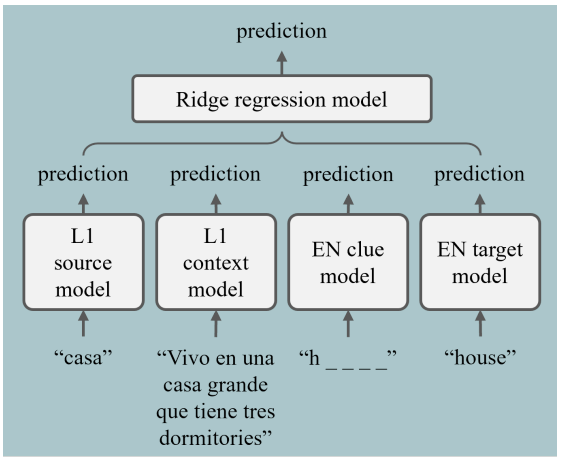
For the ensemble models, each part of the text was processed and tokenised separately.

5.1 Model architectures

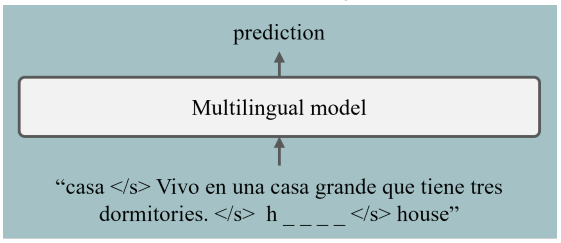
Figure 1 provides an overview of the different architectures explored for this study. For the individual monolingual models and multilingual models, the 768-dimensional embedding of the first token (<s> for RoBERTa-based models and [CLS] for BERT-based models) from the final hidden layer is passed through a dropout layer followed by a single linear layer (the regression head) to predict the difficulty score. For the monolingual ensemble models, the predictions from the individual component models are stacked together and passed through a Ridge re-



(a) Individual component models



(b) Ensemble of monolingual models



(c) Multilingual model

Figure 1: Model architectures for transformer-based approaches.

gression model. Each L1-specific ensemble learns a distinct weighting scheme for the predictions.

5.2 Model selection

Multiple pre-trained transformer models available through the Hugging Face platform[3] were considered for use in the architectures explored in this research. Preliminary model evaluation was carried out in order to select the best model for each of the architectures. Using a fixed set of hyperparameters, candidate models for each of the architectures described above were evaluated. The models were fine-tuned with the train set, and tested with the development set, reporting the RMSE and RHO of the predictions for the best model after five epochs. From this investigation, the following models were selected for further experimentation (see Table A.2 in the Appendix for results for all candidate models):

Multilingual model: XLM-RoBERTa (Conneau et al., 2020) is pre-trained on text from 100 languages using a large-scale CommonCrawl-based corpus. It employs SentencePiece tokenisation and is trained with a masked language modelling (MLM) objective. XLM-RoBERTa has been shown to outperform other multilingual transformer models in multiple NLP tasks, including cross-lingual complex word identification (Zaharia et al., 2020).

Monolingual English models: BERT (Devlin et al., 2019) is pre-trained on English text from BooksCorpus and Wikipedia using a WordPiece tokeniser. It learns contextualised word representations through masked language modelling (MLM) and next sentence prediction (NSP).

Monolingual L1 models: BERT models pre-trained for Spanish[4] (Cañete et al., 2020), German[5] (Chan et al., 2020) and Chinese[6] (Devlin et al., 2019). These models follow the BERT architecture and are pre-trained using equivalent L1 texts. For consistency, where relevant we use the cased, base model versions of each of the models listed above.

5.3 Model fine-tuning

Each of the models selected for experimentation was tuned for optimal hyperparameters. With a batch size fixed at 32 and dropout rate set to model defaults (0.1 for all models), Optuna[7], a hyperparameter optimisation framework for Python, was used to search for the best learning rate, weight decay and warm up ratio for each of the models. The models were fine-tuned with the train set, and evaluated with the development set, reporting the RMSE and RHO of the predictions for the best model after five epochs. See Table A.3 in the Appendix for the best hyperparameters for each model.

Using the optimised hyperparameters, four sets of models were fine-tuned on the train and development sets and evaluated on the test set. These included: (1) individual models for each test item component (L1 source word, L1 context, EN

[3] https://www.huggingface.co

[4] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

[5] https://huggingface.co/deepset/gbert-base

[6] https://huggingface.co/google-bert/bert-base-chinese

[7] https://optuna.readthedocs.io/en/

Model	ES		DE		CN		L1 average	
	RMSE	RHO	RMSE	RHO	RMSE	RHO	RMSE	RHO
L1 source word	0.156	0.522	0.142	0.565	0.130	0.561	0.143	0.550
L1 context	0.168	0.432	0.159	0.424	0.137	0.507	0.155	0.455
EN clue	0.169	0.400	0.155	0.389	0.139	0.477	0.154	0.422
EN target word	0.145	0.633	0.135	0.625	0.111	0.727	0.130	0.661

Table 1: RMSE and Spearman's Rho for individual component models evaluated on the KVL test set.

Model	ES		DE		CN		L1 average	
	RMSE	RHO	RMSE	RHO	RMSE	RHO	RMSE	RHO
EN target word	0.145	0.633	0.135	0.625	0.111	0.727	0.130	0.661
Monolingual ensemble	0.142	0.646	0.129*	0.651	0.106*	0.747*	0.126	0.681
Multilingual (L1-specific)	0.126*	0.734*	0.116*	0.776*	0.108	0.725	0.117	0.745
Multilingual (all-in-one)	0.116*	0.775*	0.108*	0.793	0.097*	0.785*	0.107	0.785

*Statistically significant improvement in performance compared to the prior model in the table. Significance testing was not applied to L1 average results.

Table 2: RMSE and Spearman's Rho results for the transformer-based architectures evaluated on the KVL test set.

clue, and EN target word), fine-tuned separately for each L1 subset; (2) monolingual ensembles fine-tuned per L1 subset; (3) multilingual models fine-tuned per L1 subset (L1-specific); and (4) an 'all-in-one' multilingual model fine-tuned on all L1 subsets combined.

6 Results

6.1 Model performance

Table 1 reports the individual models' performance for each test item component. The EN target word model yields the highest scores across all L1s for both RMSE and RHO, and is particularly high for the Chinese subset, with a RHO of 0.73. Overall, the next best predictor is the L1 word (average correlation: 0.55), followed by L1 context (0.46) and EN clue (0.42).

Table 2 presents results for the monolingual ensemble[8], L1-specific multilingual and all-in-one multilingual models evaluated on the KVL test set, alongside the individual EN target word model serving as a baseline. Results marked with an asterisk showed significant improvement in performance compared to the prior model in the table. On average, the ensemble architecture offers a small improvement in performance over the EN target word model for both RMSE and RHO, however the increase is not statistically significant for either metric in the ES subset, and not significant for RHO in the DE subset. The L1-specific model considerably outperforms the ensemble approach for the ES and DE subsets, with RHO increasing from 0.65 to 0.73 and 0.78, respectively. This performance increase is not seen for the CN subset, which shows a marginally poorer but non-significant performance difference for RMSE and RHO. The all-in-one model achieves the best L1 average performance in both RMSE and RHO as well as demonstrating the most consistent RHO across L1 subsets, with scores of 0.78 for ES, 0.79 for DE, and 0.79 for CN. For the DE subset, however, this performance increase is not significantly higher than the L1-specific model for RHO.

6.2 Influence of test item component

An ablation study of the test item components was conducted for the ensemble, L1-specific and all-in-one models. Single components were systematically removed from the models, in order to investigate their influence on model performance. The models were fine-tuned using the train and development set and evaluated on the test set. The full model results for RMSE and RHO coefficients with statistical significance are reported in Table A.4 in the Appendix.

Figure 2 reports the relative percentage change in RHO for each of the models after removing individual components, across the L1 subsets. Statistically significant differences in model performance are marked with an asterisk. For the monolingual ensemble models, we can see that removing the EN

[8]The learned Ridge regression weights for each component model in the ensembles can be found in Table A.1 in the Appendix.

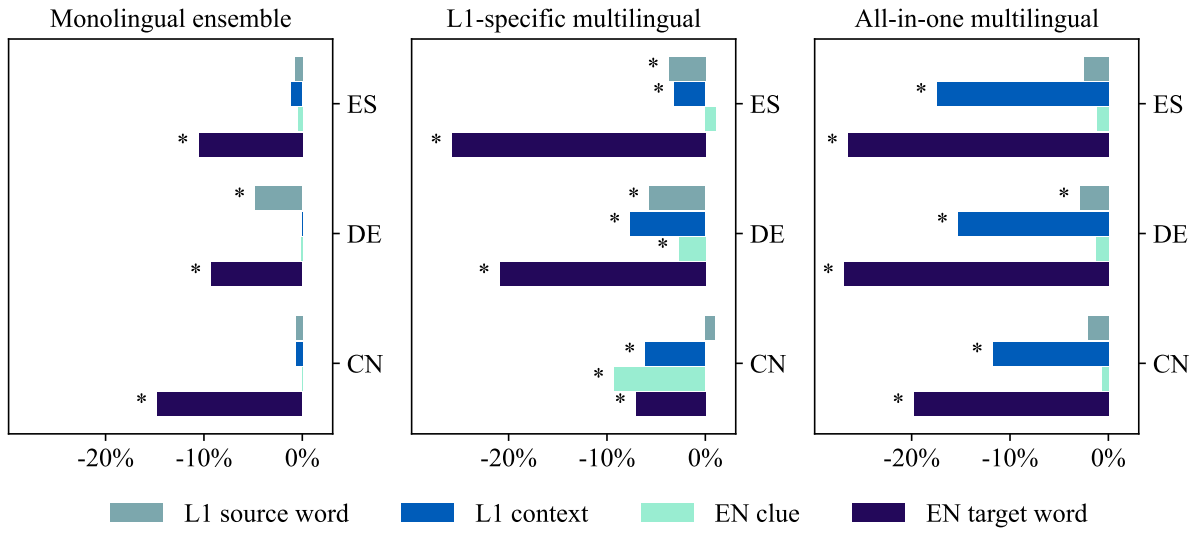


Figure 2: Relative percentage change in Spearman's Rho after individual component removal.

target word results in the largest decrease in performance – around 10% for ES and DE and 15% for the CN subset. The removal of other components have no statistically significant impact, with the exception of the DE subset which shows a small 5% degradation for the L1 word.

For the L1-specific models, we can see a more varied distribution of impact for each of the components. Statistically significant degradation of performance is seen in all three L1s for L1 context, ES and DE for L1 word and DE and CN for the EN clue. The results for the EN target word in the CN subset are notably different to those of the ES and DE, with a much lower degradation in performance after its removal (around 7%, compared to between 20-25% for DE and ES, respectively). The EN clue is in fact more impactful than the EN word in this case, showing a statistically significant 9% reduction in performance after its removal.

Looking to the all-in-one multilingual model, we can see that the ablation results begin to generalise, showing a similar pattern of impact across the L1 subsets. Results from the statistical significance tests show that removing the EN clue had no significant impact on the all-in-one model performance for any of the L1 subsets, and the removal of the L1 word has no significant impact on model performance for the ES and CN subsets.

6.3 Error analysis

Figure 3 shows the model residuals plotted against the difficulty values for each L1 subset tested on the best performing model, the all-in-one multilingual model. The graph shows that the majority of predictions fall within a range of -0.2 and +0.2 of the difficulty values. Across all L1 subsets, the model tends to underestimate the difficulty of test items as vocabulary item difficulty increases. This pattern becomes more pronounced for items with difficulty values of approximately 0.6 and above. The extent to which these higher difficulty values are impacted needs to be interpreted with caution. First of all, the GLMM difficulty estimates that the model was trained on have their own degree of error; see Schmitt et al. (2024) for further details. In addition, the fact that the GLMM scores were scaled linearly to values between 0 and 1 may also

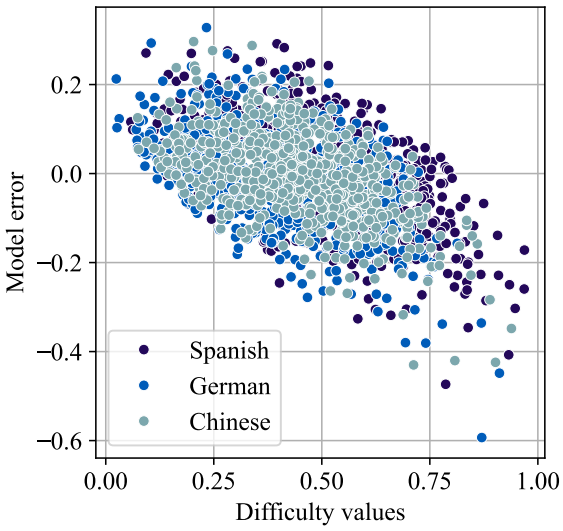


Figure 3: Error plot for the all-in-one multilingual model across L1s.

Figure 4: Example SHAP output for the ES test item for the EN target word "bar" (verb).

impact the distribution of the difficulty values at the low and high end of the scale.

In order to investigate the token-level contributions of the input text to the model predictions, further analysis using SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) was carried out. SHAP is a python package [9] that assigns Shapley values – a game-theoretic attribution metric – to features of a given predictive model. When applying SHAP to transformer architectures, each token of the input text is treated as an individual feature, affording the investigation of specific words or subword units within the sequence. In our application, this allows for fine-grained interpretability of how tokens within different components of the input text contribute to the model's final prediction.

For each L1 subset of the KVL test data, the top 10% of model errors (68 vocabulary test items per L1) were individually inspected using SHAP. For each item text, the token that contributed the most to the incorrect prediction was recorded, along with which component it was part of. Figure 4 provides an example of the SHAP analysis output for the ES item text for the English target word "bar" (verb). All tokens highlighted in red in the figure contribute to increasing the model's prediction (towards difficult) and all tokens highlighted in blue contribute to decreasing the model's prediction (towards easy). For this example, the model predicted the item to be too easy (prediction = 0.52, label = 0.93, error = -0.41). On inspecting the SHAP output, we can see that the EN target word "bar" is the token that contributes the most to the erroneous prediction.

Figure 5 shows the component and prediction direction of the tokens identified in the analysis procedure described above. Reflecting the general tendency of the model predictions reported in Figure 3, there was a higher proportion of 'too-easy' predictions (57% of errors investigated). Tokens identified in the EN target word component account for 44% of the items investigated, followed by tokens in the L1 context (25%), L1 word (14%) and EN clue (7%). The separation token <\s> was also identified as containing the top contributing token

[9] https://shap.readthedocs.io/en/latest/index.html

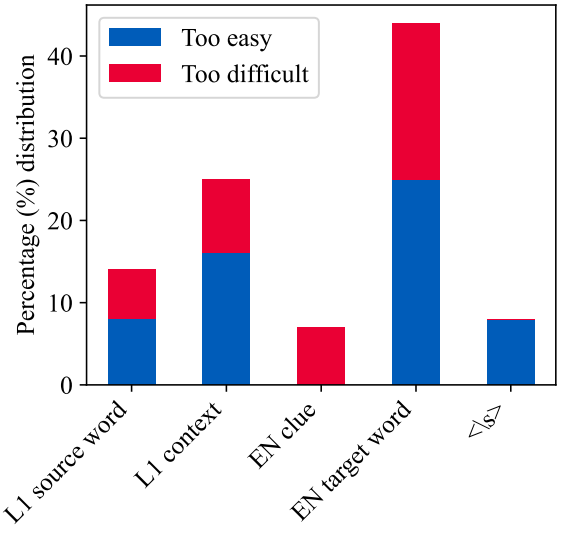


Figure 5: Location of the most influential tokens attributed to the top 10% of model errors.

for 8% of the errors (see Table A.5 in the Appendix for an overview of all identified tokens). On inspecting the tokens, some global patterns emerged.

- **Simple vs. complex words in the L1 word and L1 context.** For items that the model predicted as too easy, simple or common words were often given high attributions. For example, pronouns:"ich" (I), "我" (I), "mich" (me), or every day words: "饭" (meal), "heute" (today), "noche" (night). For items that were predicted too difficult, attribution tended to be given to more complex words such as "precisar" (specify) and "排放" (emission).

- **Sub-word tokenization in EN word.** For items that the model predicted as too easy, the sub-word with the highest attribution was often a simpler or more common word nested within the target word, for example with the compound nouns "*book*case", "*sun*shine" and "work*day*". For items that were predicted too difficult, words were often split into non-morphologically aligned subtokens: "*pois*on", "*firem*an", "ki*ller*", or suffixes: "question*able*", "punish*ment*".

- **Difficult senses in EN target word**. Whole EN target words accounted for 26% of all items that were predicted too easy, compared to only 8% for items predicted as too difficult. Common features of these EN target words were difficult senses (e.g. “short” as an adverb, “bar” as a verb), cognates with low frequency (e.g. “crystal”, “tragic”), or avoiding cognates in L1 source word (e.g. using “rastrear” instead of “explorar” for EN target word “explore”).

7 Discussion

The experiments and analysis detailed above explored L2 English vocabulary test item difficulty prediction using transformer-based architectures, with a view to establishing: (1) the best model for prediction; (2) the relative importance of test item components; and (3) potential areas of systematic bias in the best model. The outcomes of these aims are discussed below.

Initial results from fine-tuning individual component models highlighted the predictive strength of each component in isolation, with the EN target word input emerging as the most effective standalone predictor. These findings align with prior work in QDET for vocabulary testing, which has shown that features based solely on the English target word can yield a strong performance (Suyong and Hua, 2018; Ehara, 2018; Settles et al., 2020). The English target word model performed especially well on the Chinese subset, achieving a RHO of 0.73, compared to 0.63 for both Spanish and German. This may reflect the inconsistent role of cognateness in shaping word difficulty: while Spanish and German learners may be influenced by cognates or “false friends" (Otwinowska and Szewczyk, 2019), English word difficulty for Chinese learners—whose L1 shares no cognates with English—may be more directly linked to features solely attributed to the English word. As a result, the relationship between the English target word and item difficulty may be easier to model for the Chinese subset of the KVL. A similar pattern was observed by Schmitt et al. (2024) in their analysis of the KVL, where GLMM difficulty scores for the Chinese subset correlated more strongly with word frequency –a feature often used as a proxy for word difficulty (Hashimoto and Egbert, 2019)– than they did for Spanish and German.

Although the monolingual ensemble models showed limited improvement over the simpler English target word models, there may be some settings where this architecture is a suitable choice. Given the statistically significant improvement seen for the ensemble model fine-tuned with the Chinese subset, it may be that this approach is better suited to non-cognate language pairs where cross-lingual interaction does not play an important role in determining item difficulty. Furthermore, the ensemble model weights can be used as a simple proxy for component importance, offering an efficient, broader view of component relevance that may be more practically applicable for test item piloting. However, for scenarios similar to this study, in a multilingual setting with target words and context, our results suggest that a unified multilingual transformer architecture is the best choice. These findings align with prior research in multilingual LCP, which highlight the benefits of including sentence context (Bani Yaseen et al., 2021; Kelious et al., 2024a) as well as the joint modelling of different L1s (Zaharia et al., 2020).

Findings from the ablation study highlighted the advantage of cross-component representation learning within a unified transformer architecture and revealed interesting insights into the impact of fine-tuning on all L1s. Results showed that in the L1-specific approach, the model fine-tuned for Chinese assigns less importance to the English target word input compared to its Spanish and German counterparts. This is somewhat unexpected, given the very strong performance of the English target word model for Chinese shown in Table 2. One possible explanation is that the L1-specific model fine-tuned on the Chinese subset is less able to align representations of English and Chinese source words due to the lack of script overlap. This is reflected in prior research showing that multilingual models benefit from shared subword representations across languages, and that subword overlap correlates with cross-lingual transfer performance (Wu and Dredze, 2019; Pires et al., 2019). In the case of Chinese, the absence of shared subwords with English may limit the model’s ability to learn cross-lingual connections. This may be a contributing factor as to why there is no significant improvement in the L1-specific model compared to the ensemble approach for the Chinese subset.

Building on this idea, the ablation results for the all-in-one multilingual model were much more consistent across L1 subsets. The observed generalisation suggests that the all-in-one model may be learning broader, language-independent features

of vocabulary item difficulty compared to the L1-specific and ensemble models. The parallel structure of the KVL dataset, where each of the English target word and clue appears across three different L1s, likely supports this generalisation by encouraging the model to disentangle language- and item-specific features from global patterns. Furthermore, the distribution of component impact for the Chinese subset of the L1-specific model reported in Figure 2 shifts considerably toward the Spanish and German distributions seen in the all-in-one model. This may be an indication that the limitations of cross-lingual transfer for orthographically distant language pairs described above are alleviated in this setting when models are fine-tuned jointly across languages with parallel data.

Findings from the error analysis revealed valuable insights about the systematic behaviour of the all-in-one multilingual model. In addition to the effects of label re-scaling and GLMM model error discussed in Section 6.3, the normal distribution of difficulty values in the KVL dataset may further contribute to the all-in-one model's tendency to under-predict higher difficulty items. To test this, it would be of value to investigate the impact of including a larger proportion of high difficulty items during fine-tuning. This could be achieved using data-augmentation or re-sampling methods (Pan et al., 2021; Kelious et al., 2024b), or even the development of further KVL test items.

The small-scale SHAP analysis on the multilingual model's top 10% of errors, provided some general observations that can be applied to the future development of knowledge-based vocabulary lists, and test item writing more generally. In particular, the findings illustrated the impact of vocabulary complexity in the L1 word and L1 context components, suggesting that careful consideration of the word choices in the item text is needed when creating such resources for the NLP domain. Issues from the SHAP analysis that emerged relating to model behaviour, such as non-morphologically aligned sub-word tokenization and poor word sense disambiguation provide direction for improving the all-in-one model, such as multi-task learning with POS-tagging, morphological supervision or cross-lingual word sense disambiguation. Finally, given the limited scope of the SHAP-based analysis, interpretations are isolated to the individual word and subword level. Further investigation into the model's attention across tokens may be able to provide richer insight into the model behaviour.

8 Future Work

In addition to the suggestions outlined in the discussion above, there are several further avenues for future work. First, model probing for features previously found to be predictive of vocabulary item difficulty (Dunn, 2024; Hashimoto and Egbert, 2019) could help explore the item text beyond the component level, to uncover which linguistic correlates of item difficulty are being captured by the models. The all-in-one multilingual model could be further optimised by incorporating architectural adaptations shown to benefit QDET and LCP in other domains, such as scalar mixing (Gombert et al., 2024) or concatenating transformer embeddings with linguistically derived features (AlKhuzaey et al., 2024; North et al., 2023). Given the requirement of large amounts of training data for encoder-based transformer approaches, it would also be of value to compare the all-in-one model results to zero-shot and few-shot methods using LLMs, such as those recently investigated by Smădu et al. (2024). Finally, expanding the KVL dataset to include additional L1 subsets, especially those orthographically distant from English, will contribute to further exploring the role of cross-lingual transfer within multilingual transformer models, helping to corroborate the findings of this research.

9 Conclusion

This research investigated the use of transformer-based architectures for predicting vocabulary item difficulty, applying recent advances in multilingual and cross-lingual lexical complexity prediction to question difficulty estimation. Leveraging the content and structure of the KVL dataset—which has not previously been used in NLP research—this study examined the effects of multilingual text items across several transformer-based architectures. The analysis provided insights into the relative importance of different test item components across L1s, revealing how these models capture and generalise features of item difficulty. In particular, a multilingual model fine-tuned on data with all L1 variations demonstrated the strongest performance, benefiting from cross-lingual transfer and the parallel structure of the KVL dataset to produce more generalised and consistent attributions across L1s. These findings point to the potential of L1-agnostic, explainable transformer-based models for supporting test development pipelines through scalable and interpretable item calibration.

Limitations

One limitation of our study is the use of the probabilistic values derived from the GLMM framework as observed difficulty values, an issue that is discussed in more detail by Schmitt et al. (2024). To address this, we used a non-parametric correlation measure (Spearman's Rho) to evaluate our models based on rank ordering. This approach helps account for the potential error in the precision of estimates that might not be fully captured by RMSE.

Another limitation that is specific to the all-in-one multilingual model lies in the way training data was combined across L1s. The GLMM difficulty values used as labels in the models were derived from different population samples for each L1, which could raise questions about the comparability of these values across languages. To mitigate this, target labels were derived by concatenating the individually scaled subsets rather than applying a single normalisation across the entire KVL dataset. While this approach preserves the internal structure of each L1 subset difficulty scores, it does not fully account for differences in score distribution origins. However, given that predictions improved when the model was evaluated on individual L1 subsets, the all-in-one model can still be viewed as a practical means of enhancing L1-specific performance, rather than as a universal predictor of item difficulty.

Acknowledgments

This research was possible thanks to the work carried out by Norbert Schmitt (University of Nottingham, UK), Karen J. Dunn (British Council, UK), Barry O'Sullivan (British Council, UK), Laurence Anthony (Waseda Univeristy, Japan) and Benjamin Kremmel (University of Innsbruck, Austria) who created the original Knowledge-based Vocabulary Lists.

References

Samah AlKhuzaey, Floriana Grasso, Terry R Payne, and Valentina Tamma. 2024. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.

Tuqa Bani Yaseen, Qusai Ismail, Sarah Al-Omari, Eslam Al-Sobh, and Malak Abdullah. 2021. JUST-BLUE at SemEval-2021 task 1: Predicting lexical complexity using BERT and RoBERTa pre-trained language models. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 661–666, Online. Association for Computational Linguistics.

Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2023. A survey on recent approaches to question difficulty estimation from text. *ACM Comput. Surv.*, 55(9).

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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, Online. Association for Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Bradley Efron. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185.

Yo Ehara. 2018. Building an English vocabulary knowledge dataset of Japanese English-as-a-second-language learners using crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Pierre Finnimore, Elisabeth Fritzsch, Daniel King, Alison Sneyd, Aneeq Ur Rehman, Fernando Alva-Manchego, and Andreas Vlachos. 2019. Strong baselines for complex word identification across multiple

languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 970–977, Minneapolis, Minnesota. Association for Computational Linguistics.

Sebastian Gombert, Lukas Menzel, Daniele Di Mitri, and Hendrik Drachsler. 2024. Predicting item difficulty and item response time with scalar-mixed transformer encoder models and rational network regression heads. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 483–492, Mexico City, Mexico. Association for Computational Linguistics.

Brett J Hashimoto and Jesse Egbert. 2019. More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning*, 69(4):839–872.

Abdelhak Kelious, Mathieu Constant, and Christophe Coeur. 2024a. Complex word identification: A comparative study between ChatGPT and a dedicated model for this task. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3645–3653, Torino, Italia. ELRA and ICCL.

Abdelhak Kelious, Mathieu Constant, and Christophe Coeur. 2024b. Investigating strategies for lexical complexity prediction in a multilingual setting using generative language models and supervised approaches. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 96–114, Rennes, France. LiU Electronic Press.

Yoolim Kim, Vita V. Kogan, and Cong Zhang. 2024. Collecting big data through citizen science: Gamification and game-based approaches to data collection in applied linguistics. *Applied Linguistics*, 45(1):198–205. Published: 12 July 2023.

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

Ekaterina Loginova, Luca Benedetto, Dries Benoit, and Paolo Cremonesi. 2021. Towards the application of calibrated transformers to the unsupervised estimation of question difficulty from text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 846–855, Held Online. INCOMA Ltd.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30.

Kiyoaki Nakanishi, Nobuyuki Kobayashi, Hiromitsu Shiina, and Fumio Kitagawa. 2012. Estimating word difficulty using semantic descriptions in dictionaries and web data. In *2012 IIAI International Conference on Advanced Applied Informatics*, pages 324–329.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Comput. Surv.*, 55(9).

Agnieszka Otwinowska and Jakub M. Szewczyk. 2019. 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.

Chunguang Pan, Bingyan Song, Shengguang Wang, and Zhipeng Luo. 2021. DeepBlueAI at SemEval-2021 task 1: Lexical complexity prediction with a deep ensemble approach. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 578–584, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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

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

Kim Cheng Sheang. 2019. Multilingual complex word identification: Convolutional neural networks with morphological and linguistic features. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 83–89, Varna, Bulgaria. INCOMA Ltd.

Răzvan-Alexandru Smădu, David-Gabriel Ion, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. 2024. Investigating large language models for complex word identification in multilingual and multidomain setups. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16764–16800, Miami, Florida, USA. Association for Computational Linguistics.

Yuni Susanti, Takenobu Tokunaga, and Hitoshi Nishikawa. 2020. Integrating automatic question generation with computerised adaptive test. *Research and Practice in Technology Enhanced Learning*, 15:1–22.

Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. 2017. Controlling item difficulty for automatic vocabulary question generation. *Research and practice in technology enhanced learning*, 12:1–16.

Eum Suyong and Yang Hua. 2018. Feature analysis on English word difficulty by gaussian mixture model. *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 191–194.

Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2024. Question difficulty prediction based on virtual test-takers and item response theory. In *Workshop on Automatic Evaluation of Learning and Assessment Content (EvalLAC'24)*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Kevin P. Yancey, Andrew Runge, Geoffrey LaFlair, and Phoebe Mulcaire. 2024. BERT-IRT: Accelerating item piloting with BERT embeddings and explainable IRT models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 428–438, Mexico City, Mexico. Association for Computational Linguistics.

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482, Mexico City, Mexico. Association for Computational Linguistics.

George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020. Cross-lingual transfer learning for complex word identification. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 384–390.

Leonidas Zotos, Hedderik van Rijn, and Malvina Nissim. 2025. Can model uncertainty function as a proxy for multiple-choice question item difficulty? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11304–11316, Abu Dhabi, UAE. Association for Computational Linguistics.

A Appendix

Component model	ES	DE	CN
L1 source word	14.97%	39.93%	23.93%
L1 context	18.80%	6.46%	13.94%
EN clue	5.69%	1.66%	0.00%
EN target word	60.54%	51.93%	62.13%

Table A.1: Ridge regression ensemble model weights across L1 subsets.

Input text	Pre-trained model	ES		DE		CN		L1 average	
		RMSE	Corr	RMSE	Corr	RMSE	Corr	RMSE	Corr
L1 context	BERT-mono*	0.139	0.449	0.133	0.437	0.125	0.503	**0.132**	**0.463**
L1 context	XLM-R	0.140	0.429	0.134	0.429	0.127	0.483	0.134	0.447
L1 context	mBERT	0.141	0.416	0.135	0.412	0.128	0.451	0.135	0.426
L1 source word	BERT-mono*	0.135	0.496	0.123	0.584	0.118	0.596	**0.125**	**0.559**
L1 source word	XLM-R	0.149	0.376	0.132	0.486	0.119	0.555	0.133	0.472
L1 source word	mBERT	0.142	0.409	0.129	0.506	0.119	0.543	0.130	0.486
EN clue	BERT	0.144	0.365	0.139	0.358	0.131	0.396	**0.138**	**0.373**
EN clue	RoBERTa	0.145	0.355	0.139	0.343	0.131	0.392	**0.138**	0.363
EN target word	BERT	0.123	0.592	0.116	0.629	0.096	0.752	**0.112**	**0.658**
EN target word	RoBERTa	0.134	0.513	0.133	0.506	0.110	0.624	0.126	0.548
All components	XLM-R	0.103	0.761	0.099	0.777	0.088	0.800	**0.097**	**0.779**
All components	mBERT	0.107	0.744	0.103	0.751	0.094	0.773	0.101	0.756

*BERT-mono refers to the monolingual models for each L1 outlined in Section 5.2. Hyperparameters were fixed at 2e-5 for learning rate, 0.1 for weight decay and 0.1 for warm up ratio.

Table A.2: RMSE and Spearman's Rho for each of the transformer models considered for the final experiments. Models were fine-tuned on the train set and evaluated on the development set.

Model name	Input language	Input text	Learning rate	Weight decay	Warmup ratio
bert-base-spanish-wwm-cased	ES	L1 source word	3e-5	0	0.1
gbert-base	DE	L1 source word	2e-5	0	0.1
bert-base-chinese	CN	L1 source word	2e-5	0.1	0
bert-base-spanish-wwm-cased	ES	L1 context	3e-5	0	0.1
gbert-base	DE	L1 context	2e-5	0	0.1
bert-base-chinese	CN	L1 context	3e-5	0	0
bert-base-cased	ES	EN clue	2e-5	0.1	0.1
bert-base-cased	DE	EN clue	2e-5	0.1	0
bert-base-cased	CN	EN clue	3e-5	0	0
bert-base-cased	ES	EN target word	3e-5	0	0
bert-base-cased	DE	EN target word	1e-5	0	0.1
bert-base-cased	CN	EN target word	2e-5	0	0
xlm-roberta-base	ES	All components	3e-5	0.1	0.1
xlm-roberta-base	DE	All components	3e-5	0	0.1
xlm-roberta-base	CN	All components	3e-5	0.1	0.1
xlm-roberta-base	XX	All components	3e-5	0.1	0.1

Search space for hyperparameters: learning rate (1e-5, 2e-5, 3e-5), weight decay (0, 0.1), warm up ratio (0, 0.1).

Table A.3: Optuna hyperparameter results for the models selected for the final experimentation. Models were fine-tuned on the train set and evaluated on the development set.

Full component model	**ES**		**DE**		**CN**		**L1 average**	
- removed component	RMSE	RHO	RMSE	RHO	RMSE	RHO	RMSE	RHO
Ensemble	0.142	0.646	0.129	0.651	0.106	0.747	0.126	0.681
- L1 word	0.142	0.641	0.134*	0.620*	0.108	0.742	0.128	0.668
- L1 context	0.142	0.639	0.128	0.651	0.107	0.742	0.126	0.677
- EN clue	0.142	0.643	0.129	0.650	0.106	0.747	0.126	0.680
- EN word	0.155*	0.578*	0.138*	0.591*	0.122*	0.637*	0.138	0.602
L1-specific	0.126	0.734	0.116	0.776	0.108	0.725	0.117	0.745
- L1 word	0.133*	0.707*	0.122*	0.732*	0.110	0.732	0.122	0.724
- L1 context	0.129	0.711*	0.123*	0.717*	0.116*	0.681*	0.123	0.703
- EN clue	0.125	0.742	0.119	0.755*	0.119*	0.658*	0.121	0.718
- EN word	0.152*	0.545*	0.138*	0.614*	0.117*	0.674*	0.136	0.611
All-in-one	0.116	0.775	0.108	0.793	0.097	0.785	0.107	0.784
- L1 word	0.122*	0.756	0.112	0.770*	0.102*	0.769	0.112	0.765
- L1 context	0.138*	0.640*	0.126*	0.672*	0.114*	0.693*	0.126	0.668
- EN clue	0.121*	0.766	0.115*	0.783	0.101*	0.780	0.112	0.776
- EN word	0.151*	0.570*	0.139*	0.580*	0.123*	0.630*	0.138	0.593

*Statistically significant improvement in performance compared to the 'full component' models (as reported in Table 2). Significance testing was not applied to L1 average results.

Table A.4: RMSE and Spearman's Rho results for the ablation study models. Models were fine-tuned on the train and development set, and evaluated on the test set.

	Too easy			Too difficult		
	ES	DE	CN	ES	DE	CN
L1 source word	para**milit**ares doble **calcula**dora **anal**íticas **restaur**ada ol**ím**pica	falsch **Mili**tia **Physik**er **ir**anisch **Or**chester	找 账 笑 坐 的人	precisar e**mbaj**ada discusión	(denk**würdig** novice her**ab**fallend **Zu**hörer	资 属 重大 (
L1 context	coche Mi **domingo**s noche **mezcla**r **pa**vor físico **verdur**as comida	**Imit**ator heute **Un**recht Front ich Beunruhig**un**g wird wirklich mich Handy und Ich	我 旅行 音 我 骄傲 去 饭 信任 从小 回家 人们 我	bomba en**tero** media **imperio**sa la investigacion profesional efectivo	künstler**ische**	涌 的看法 美洲 腺 系统 器 养 郊 排放
EN clue*	-	-	-	m_____ (masterpiece) n_____ (nominated) n____ (nurse) e_____ (expensive) t_____ (terrorism)	b_____ (better) n____ (nurse) t_____ (traffic) q___ (quit) p___ (plus) m_____ (masterpiece) r_____ (reality)	o_____ (olympic) o_____ (oversize) e_____ (examination)
EN target word	**un**ity bar grand **ja**smine tragic glo**w** grin **sun**shine explore pasta domina**te** **introduc**tion lyrics **re**cycled **trop**ics re**cycling**	short bar **communica**tor grand forget crystal tragic cheer**ful** vote kangaro**o** learned boost **domina**ted recipe eco**system** sitting café **book**case	bar **un**ity grand short **book**case stop tried glo**w** written **che**erful **go**wn work**day** recipe learned birth tragic cite sweat	**pois**on **quanti**ty **memor**able excellent **fall**ing venture **incomp**lete question**able** **firem**an **cho**rus **in**coming	kidnap**ping** **face**less climate question**able** in**com**ing bala**ncing** **gu**ilt **definit**e **minori**ty established **break**out **pois**on taking **climb**ing	chi**nese** **relat**e **re**ly ki**ller** government**al** inexpen**sive** disorder punish**ment** question**able** quali**fy** back**ward** **anti**social issue **firem**an

*The associated EN target word for the EN clue component is included in brackets for interpretability. Within the component groups, words are listed in order of largest to smallest model prediction error for their associated item.

Table A.5: The words and subwords (in bold) contributing to the top 10% of the all-in-one multilingual model errors, according to SHAP analysis.