Learning Shortcuts: Task-Level Confounding in Automatic Leveling of ELE Student Texts

Bernardo Sigüenza-Moreno

Independent Researcher. Huelva, España

Abstract

This paper presents our participation in the NivELE shared task within the IberLEF 2026 evaluation campaign, which consists of assigning a CEFR level to texts written by learners of Spanish as a Foreign Language (ELE). The submitted system, based on fine-tuning BSC-LT/MrBERT-es on a fused corpus (CEDEL2, CORANE, and an unverified HuggingFace scrape of CAES), achieved an F1-macro of 0.60279 on the official test set, ranking first on the NivELE 2026 leaderboard. This score could not be reproduced in subsequent training runs with identical hyperparameters. The stable behavior of the system, estimated from a 5-fold cross-validation reproducibility analysis, corresponds to a mean leaderboard F1-macro of 0.5531 (SD = 0.0489); all analyses in this paper are based on these reproducible results, not on the non-reproducible submission score. The reproducibility analysis revealed a mean internal validation score of 0.9172 against a mean leaderboard score of 0.5531, a difference of 0.3641 points. This gap is consistent with the hypothesis that the system learned thematic rather than linguistic proficiency signals, exploiting task-level confounding variable present in the training corpus. This hypothesis, however, cannot be demonstrated: the absence of task and level metadata in the official test set precludes any quantification of confounding.

Keywords

CEFR Level Assessment, Automatic Essay Scoring, Confounding Variables, Transformer-based models

1. Introduction

The initial placement of ELE students is costly in terms of teaching time, since each student's test must be assessed individually using established methodological criteria to support their appropriate progression. NivELE [1], within the IberLEF 2026 [2], stands as the first shared task to address automatic proficiency-level assignment for ELE.

In this paper, we present our participation in the NivELE task. To examine the feasibility of automating the assessment of written texts using Natural Language Processing techniques, we specifically explore ModernBERT as an ELE-level classifier.

The rest of the paper is structured as follows: Section 2 reviews related works, Section 3 describes the task, Section 4 presents the training dataset and the system developed, while the results regarding this system are gathered in Section 5. Section 6 analyzes the limitations of the research arising from the test corpus. Conclusions close the paper.

2. Related Works

The application of Transformer-based language models to the automatic assessment of linguistic proficiency in foreign language learners has been extensively explored in works that investigate encoder-only architectures (BERT) [3] and decoder-only architectures (GPT) [4].

Without claiming to be exhaustive, we review several prior works that addressed the task of CEFR-level classification across different languages using both of the aforementioned approaches:

- Encoder-only: Schmaltz and Brutti [5] apply BERT to EFCAMDAT [6] and CLC-FCE [7] to classify English learner texts into CEFR levels, exploring input augmentation with error corrections. Rama

IberLEF 2026, September 2026, León, Spain

besiguenza@hotmail.com (B. Sigüenza-Moreno)

0009-0006-7772-253X (B. Sigüenza-Moreno)

and Vajjala [8] explore the utility of multilingual pretrained representations for CEFR classification. Jalota et al. [9] investigate strategies to mitigate the influence of L1 on learner texts. Borah et al. [10] address the problem of spurious correlations in neural translationese classifiers. Using unsupervised topic modeling (LDA and BERTopic), they formalize an alignment measure between topics and class labels, which they show to be equivalent to purity in clustering, and propose the concept of a "topic floor": an upper bound on what a classifier may learn from spurious topical correlations rather than genuine signals of the target phenomenon.

- Decoder-only: Mizumoto and Eguchi [11], Yancey et al. [12], Banno et al. [13] and Cantero-Romero et al. [14] investigate the capacity of different GPT versions to assess the written proficiency of L2 learners. Mizumoto and Eguchi [11] evaluate the accuracy of GPT-3 on automated essay scoring (AES) using 12,100 texts from the TOEFL11 corpus [15], reporting performance gains when linguistic features are incorporated. Yancey et al. [12] employ GPT-3.5 and GPT-4 to score texts written by English L2 learners, measuring agreement with human raters and finding systematic discrepancies between GPT-4 and human evaluators that vary as a function of the learners' L1. Banno et al. [13] analyze the capacity of GPT-4 to explain learner linguistic proficiency, investigating whether GPT-4 can decompose L2 writing assessment into analytic subscores corresponding to the components of the CEFR. Cantero-Romero et al. [14] use GPT-3.5-turbo to compare two automated written proficiency classification systems for ELE: (i) a zero-shot prompting system, (ii) a system fine-tuned on 6,571 samples spanning levels A1-C1 from 11 L1 backgrounds drawn from CAES v2.1 [16]. Evaluation is conducted on an in-house corpus of 316 texts collected at a university language center. A qualitative error analysis revealed that the model occasionally assigns proficiency levels based on the text topic rather than on linguistic proficiency markers.

3. Task Description

The NivELE 2026 shared task focuses on the development and evaluation of natural language processing models for the automatic classification of texts written by learners of ELE according to CEFR proficiency levels. The five classes established by the organizing committee correspond to the following CEFR levels: A1, A2, B1, B2, C1.

At first glance, NivELE appears to be a standard classification task. Given an input text, the model assigns a class label (level) from the set A1, A2, B1, B2, C1. However, the nature of the task has distinctive characteristics that complicate framing the solution as a classification problem: (i) the object of study is a learned language, not a native one. The texts are productions by ELE learners, which means they contain errors, first-language interference, avoided structures, and interlanguage features; (ii) level assignment is holistic. CEFR levels are not discrete categories. The CEFR scale defines profiles of overall communicative competence, not isolated linguistic features. The level emerges from the interaction among multiple linguistic dimensions; boundaries between adjacent levels are fuzzy.

Therefore, although it is formally a classification problem, the signal that a model must learn in order to replicate a rater's judgment is deeply linguistic and multidimensional in nature.

3.1. Dataset

The NivELE 2026 shared task does not establish an official development dataset; instead, it recommends a set of available corpora and provides a sample of five examples, one per level. Each example consists of a representative text and its associated level label according to CEFR criteria.

3.2. Evaluation Metric

The primary evaluation metric for the task is macro-averaged F1 over the five classes.

4. Methodology

4.1. Dataset

The supervised approach adopted in this work requires a training dataset. This dataset was constructed from two established corpora: CEDEL2 [17] and CORANE [18], and a HuggingFace-hosted CAES scrape [19] without verification against the original CAES [16]. The resulting dataset comprises 25,076 instances with the structure (id, text, level, task, L1, number of words).

The dataset inherits the limitations of the corpora from which it derives:

- An almost structurally fixed association between text type and CEFR level. Each task type, particularly in the CAES scrape, is associated by curriculum design with one or two specific levels. This phenomenon introduces a confounding variable between task and level that directly affects any supervised use of the corpus for automatic classification.
- A noisy approximation to CEFR levels. In order to map the proprietary scales of CEDEL2 and CORANE onto the CEFR scale, the assumption is made that those scales are congruent with the CEFR scale level by level. Even if this holds, the level recorded in both corpora characterizes not the text itself but the student at the time of beginning the course. The labels of both corpora are treated as the gold standard; in reality, they constitute a noisy approximation.

We verified that the training corpus and the NivELE test corpus share no exact texts. However, 286 texts in the test corpus (316 texts) share a semantic similarity greater than or equal to 0.857 with texts in the training corpus. The 0.857 threshold is justified based on the comparison of the thematic spaces of the training and test corpora; a coverage analysis was carried out using multilingual-e5-large embeddings [20], a model trained with a contrastive semantic-similarity objective and without fine-tuning on the CEFR classification task. Figure 1 shows UMAP projections (n_neighbors 15, min_dist 0.1, metric "cosine", seed 42) [21] of the thematic spaces of both corpora. Cross-level mixing is visible. Most of the test triangles fall within, or immediately on the periphery of, regions populated by training examples. There is no zone of the thematic space where test points accumulate in the total absence of training points. This indicates that the thematic space of the test corpus is globally covered by the training corpus.

The range of cosine distances from test corpus instances to their nearest training neighbor lies in [0.05, 0.17]. The P90 is 0.1425, which implies that 90% of the test texts have a neighbor in the training set with a cosine similarity above 0.857 (1 - 0.1425). Figure 2 shows the histogram of cosine distances between test corpus instances and their nearest neighbor in the training corpus.

The statistics for the corpus training are shown in Appendix A.

4.2. Internal baselines

In order to quantify the biases imposed by the dataset, an experiment was conducted in which three SVM classifiers with an RBF kernel were trained via 5-fold cross-validation. For each fold, the dataset was partitioned into two splits (80% training, 20% validation) using a stratified random strategy that preserved the original class proportions:

- **Length.** An SVM classifier with an RBF kernel predicting the class label on the train dataset (validation split) using only text length as a predictor achieves F1-macro = 0.417 ± 0.067.
- **L1.** An SVM classifier with an RBF kernel predicting the class label on the train dataset (validation split) using only L1 as a predictor achieves F1-macro = 0.194 ± 0.012.
- **Task.** An SVM classifier with an RBF kernel predicting the class label on the train dataset (validation split) using only the task label as a predictor achieves F1-macro = 0.531 ± 0.112.

It is plausible that text length correlates with level either because writing tasks at higher levels tend to produce longer texts, or because advanced learners write more. The SVM learns length, not linguistic competence. L1 does not exceed the score that a random classifier (assigning labels with probabilities

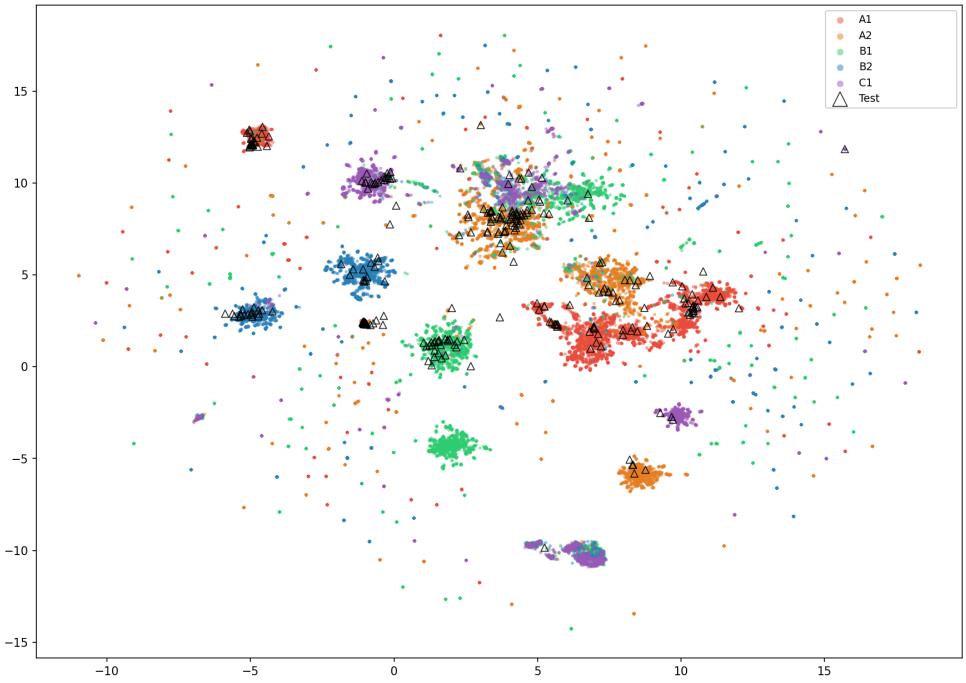

Figure 1: UMAP projections of multilingual-E5-large embeddings (without fine-tuning)

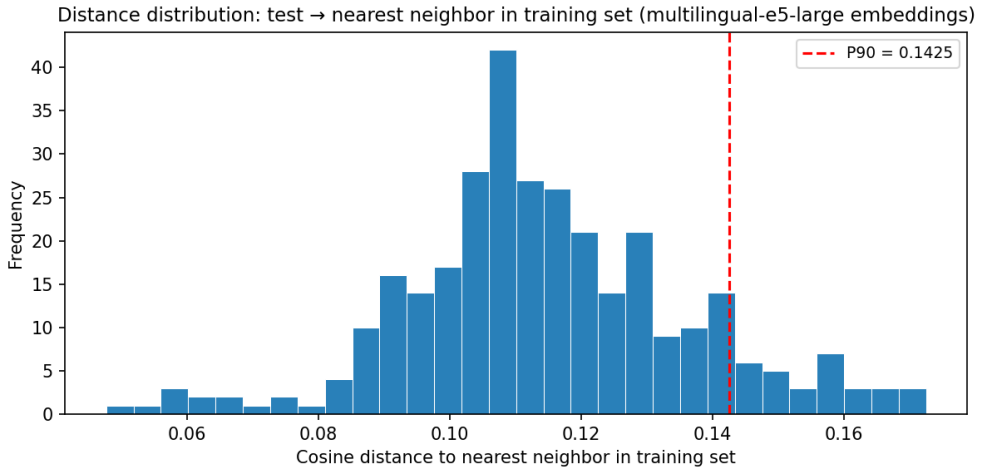

Figure 2: Histogram of cosine distances between test corpus instances and their nearest neighbor in the training corpus

proportional to the class distribution of the dataset) would achieve. Using only the task label, the model achieves an F1-macro of 0.531, which is 2.65 times above chance (0.531 / 0.200). The operational implication is that any model trained on the dataset will learn this confounding variable.

4.3. Proposed System

The proposed system uses a BSC-LT/MrBERT-es model [22], a ModernBERT variant [23] designed to process texts in Spanish and English. The choice of the encoder-only architecture is justified on the following grounds (paraphrased from the original [23]): encoder-only transformer models, such as BERT, offer an excellent performance-to-size ratio for retrieval and classification tasks, outperforming larger decoder-only models. ModernBERT incorporates modern optimizations into encoder-only models

and represents a significant Pareto improvement over previous encoders. ModernBERT models achieve state-of-the-art results across a wide range of benchmarks. In addition to strong downstream task performance, ModernBERT is also the fastest and most memory-efficient encoder and is designed for standard GPU inference.

Automatic classification of texts produced by ELE students is framed as a multiclass classification problem. The goal of the developed model is to correctly predict the class label (proficiency level) of each text. Note that the model treats each class independently, without taking into account the curricular ordering of CEFR levels and disregarding any of the other observations made about the task in the Task Description section.

The model that achieved F1-macro = 0.60279 was fine-tuned for 5 epochs with batch size: 32, learning rate: 1e-4, warmup ratio: 0.1, weight decay: 0.01, maximum input length: 1024. With a maximum input length of 1024 tokens, 18 texts (0.1%) in the training corpus are truncated (A1: 0, A2: 0, B1: 2, B2: 1, C1: 15).

A learning rate of 1e-4 exceeds the range typically recommended for fine-tuning BERT-family models. The official submission score of 0.60279 corresponds to a non-reproducible training run. Subsequent runs with identical hyperparameters did not achieve comparable F1-macro values.

Result reproducibility was assessed by constructing a new system (S1) using 5-fold cross-validation on dataset partitions (80% train / 20% validation) generated at each fold using a level-stratified random sampling strategy, preserving the original class proportions across levels. At each fold, a BSC-LT/MrBERT-es model was trained for 5 epochs with batch size 32, learning rate 1e-4, warmup ratio 0.1, weight decay 0.01, maximum input length 1024, and seed 42.

Each fold achieved the following F1-macro scores:

Fold	Validation (F1-macro)	Test (F1-macro)
1	0.9211	0.59387
2	0.9131	0.56133
3	0.9163	0.46868
4	0.9146	0.57721
5	0.9207	0.56443
Mean	0.9172	0.55310
Std	0.0032	0.04890

Table 1
S1 cross-validation results. Validation scores correspond to internal fold evaluation; test scores correspond to individual submissions to the Kaggle leaderboard.

Mean F1-macro difference (validation vs. test): 0.3641

A second system (S2) was trained using the same strategy (5-fold cross-validation, 80/20 split), modifying only the learning rate to 2e-5 and leaving all remaining hyperparameters unchanged.

Fold	Validation (F1-macro)	Test (F1-macro)
1	0.8757	0.56039
2	0.8564	0.43545
3	0.8648	0.39289
4	0.8581	0.42895
5	0.8594	0.41675
Mean	0.8629	0.44690
Std	0.0070	0.06550

Table 2
S2 cross-validation results. Validation scores correspond to internal fold evaluation; test scores correspond to individual submissions to the Kaggle leaderboard.

Mean F1-macro difference (validation vs. test): 0.4160

Finally, two consecutive fine-tuning runs (S3, S4) were performed on the full dataset, training on the dataset without held-out validation, retaining the following hyperparameters: learning rate 1e-4, batch size 32, maximum input length 1024, warmup ratio 0.1, weight decay 0.01, seed 42, epochs 5. The runs achieved F1-macro scores of 0.55543 and 0.56464 respectively, with a difference of 0.009 points between runs.

5. Discussion

In internal validation, system S1 (LR=1e-4) achieved a mean F1-macro of 0.9172 (SD=0.0036) and a mean F1-macro of 0.5531 (SD=0.0489) on the test set, corresponding to a mean difference of 0.3641 points. System S2 (LR=2e-5) showed an analogous gap: 0.8629 (SD=0.0078) in internal validation versus 0.4469 (SD=0.0655) on the test set, with a mean difference of 0.4160 points. Internal evaluation thus overestimates the actual system performance in both cases, indicating that internal validation sets and the official test set are not interchangeable as estimators of system performance.

The results achieved by S3 (F1-macro = 0.55543) and S4 (F1-macro = 0.56464), with a difference of 0.009 points between runs, document two distinct findings:

1. Non-determinism in the fine-tuning process. With a fixed seed (42), two consecutive runs on the same data do not produce the same model. This phenomenom is documented in the literature [24], [25].
2. The official system result (F1-macro = 0.60279) is not reproducible with the documented hyperparameters. The mean of the two replication runs (S3, S4) is 0.560, representing a difference of 0.043 points from the official result. This gap exceeds the variance observed between the two replications (0.009 points). The most plausible explanation is that some undocumented difference was inadvertently introduced in the training dataset or in the hyperparameters of the original run. We consider it less likely that the original run found an exceptionally favorable optimization trajectory that the training process does not reproduce reliably. Neither assumption has been verified, and both remain as limitations of this work.

An unexpected result of the comparison between S1 and S2 is that S1, whose learning rate of 1e-4 exceeds the recommended range for fine-tuning BERT-family models, outperforms S2 across all five leaderboard folds, with differences ranging from 0.034 to 0.148 F1-macro points. S1 also outperforms S2 on internal validation (0.9172 vs. 0.8629). No analysis of the causes of this result has been conducted.

It is further observed that the four stable folds of S1 on the leaderboard (trained on 80% of the dataset) yield a mean F1-macro of 0.574, which is higher than the mean of the two replication runs (S3, S4) trained on the full dataset (0.560). This reverses the expectation that a larger amount of training data produces better performance on the test set. No ablation studies were conducted to determine whether the specific composition of each partition accounts for this finding.

Fold 3 shows a marked performance drop in both systems (S1, S2) with the same seed. S1 achieves 0.469 on that fold against a mean of 0.574 across the four remaining folds; S2 achieves 0.393 against a mean of 0.460. Since the seed is identical and the learning rate is the only variable that differs between systems, the performance drop cannot be attributed to the learning rate. A plausible but unverified explanation is that it results from the composition of the training split in the third fold.

The model does not learn descriptors of linguistic competence; it learns the assignment of tasks by level. The UMAP projection (n_neighbors 15, min_dist 0.1, metric “cosine”, seed 42) on the S4 CLS embeddings of the training corpus reveals that each CEFR level appears fragmented into multiple discrete subclusters; they do not form a single region (Figure 3). In a model that had learned linguistic competence as its signal, each level should form a compact region. Adjacent levels should share overlapping boundaries in a continuous A1-C1 space. One explanation consistent with the distribution is that the model has learned to separate task types or textual genres.

The projection of the CLS embeddings of the test set onto the UMAP space of the S4 model shows that the test texts do not occupy empty regions of the representation space, which rules out severe

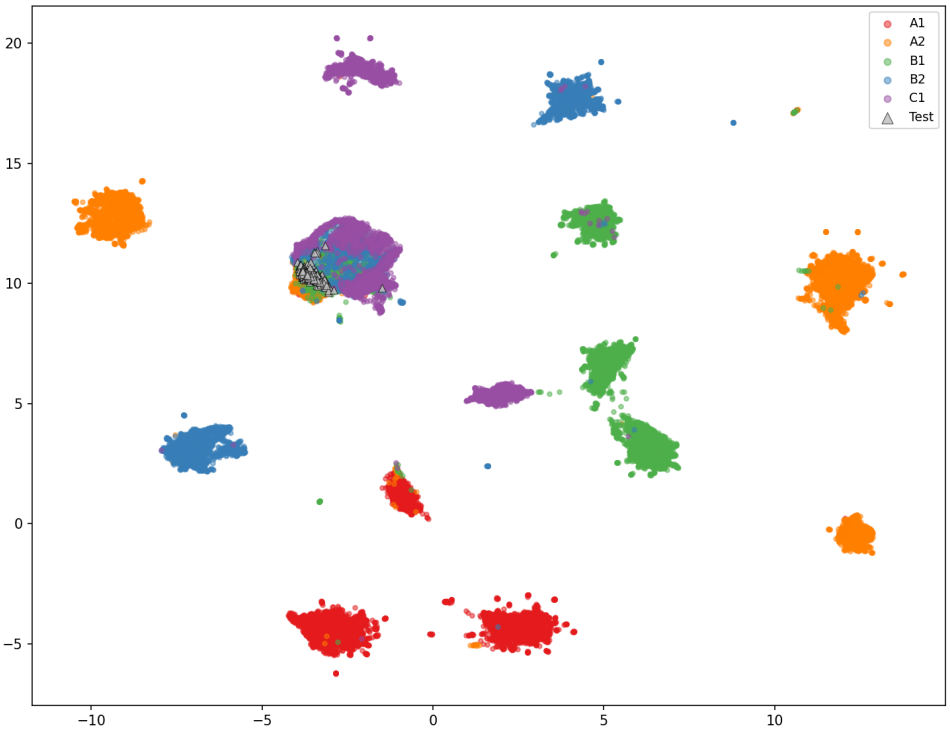

Figure 3: UMAP projections of CLS embeddings from the S4 model

distributional shift at the embedding level. The concentration of test points within a subset of training subclusters reinforces the hypothesis that the texts in the test corpus share a task or topic with texts in the training corpus.

6. Limitations

Internal evaluation presents a circularity problem that limits the interpretation of results. Both systems (S1, S2) were trained and evaluated on the same dataset, using different train and validation partitions to prevent data leakage. Since this dataset contains documented task-level confounding variable, internal evaluation cannot determine whether the model learned genuine linguistic proficiency signals or surface-level correlates of CEFR level. The official test set, accessed through the leaderboard, is the only external evaluator available, but it lacks task, L1, and proficiency level metadata that would allow diagnosis of error sources.

7. Conclusions

This paper presents our participation in the NivELE shared task within the IberLEF 2026 evaluation campaign. A single training run achieved an F1-macro of 0.60279 on the official test set, ranking first on the NivELE 2026 leaderboard; this score could not be reproduced in subsequent runs with identical hyperparameters. The stable behavior of the system, estimated from a 5-fold cross-validation reproducibility analysis, corresponds to a mean leaderboard F1-macro of 0.5531 (SD = 0.0489). All analyses and conclusions in this paper are based on this reproducible estimate. The systems developed are naive in a specific sense: they assume the model will learn linguistic features from raw text without any explicit linguistic signal in the training input. The gap of 0.3641 points between mean internal validation performance (0.9172) and mean leaderboard score (0.5531) raises the hypothesis that the classifier learned thematic rather than linguistic signals, exploiting task-level confounding variable

structurally present in the training corpus. The thematic coverage analysis shows that 90% of the test texts have a nearest neighbor in the training corpus with cosine similarity above 0.857. The hypothesis, however, cannot be demonstrated: the absence of task and L1 metadata in the official test set precludes any post hoc quantification of confounding, and the circularity of internal evaluation prevents distinguishing genuine proficiency signals from surface-level correlates. Future work should evaluate systems on corpora with controlled task and L1 distributions, and explore systems that incorporate explicit linguistic features as part of the training signal, rather than relying on raw text.

8. Declaration on Generative AI

During the development of this work, the model Claude Sonnet 4.6 (Anthropic) was used as a supporting tool for:

1. consultation on methodological questions;
2. generation of Python code fragments in an AI-assisted programming mode;
3. provisional translation of textual fragments from Spanish to English, subject to subsequent editorial review.

In all cases, responsibility for the content, empirical results, interpretations, and claims of the paper rests exclusively with the author.

References

[1] M. V. Cantero Romero, J. Collado Montañez, I. Cabrera de Castro, J. Fruns Giménez, A. Arjonilla Sampedro, J. Cruz, A. Montejo-Ráez, S. M. Jiménez-Zafra, Resumen de NivELE en IberLEF 2026: Nivelación automática de textos de estudiantes de ELE (Español como Lengua Extranjera), Procesamiento del Lenguaje Natural 77 (2026).

[2] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS. org, 2026.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), 2019, pp. 4171–4186.

[4] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

[5] V. J. Schmalz, A. Brutti, Automatic assessment of english cefr levels using bert embeddings, in: Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021), 2021, pp. 295–301.

[6] J. Geertzen, T. Alexopoulou, A. Korhonen, et al., Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat), in: Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project, 2013, pp. 240–254.

[7] H. Yannakoudakis, T. Briscoe, B. Medlock, A new dataset and method for automatically grading ESOL texts, in: D. Lin, Y. Matsumoto, R. Mihalcea (Eds.), Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 180–189. URL: https://aclanthology.org/P11-1019/.

[8] T. Rama, S. Vajjala, Are pre-trained text representations useful for multilingual and multi-dimensional language proficiency modeling?, arXiv preprint arXiv:2102.12971 (2021).

[9] R. Jalota, P. Bourgonje, J. Van Sas, H. Huang, Mitigating learnerese effects for cefr classification, in: Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), 2022, pp. 14–21.
[10] A. Borah, D. Pylypenko, C. Espana-Bonet, J. van Genabith, Measuring spurious correlation in classification:"clever hans" in translationese, in: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, 2023, pp. 196–206.
[11] A. Mizumoto, M. Eguchi, Exploring the potential of using an ai language model for automated essay scoring, Research Methods in Applied Linguistics 2 (2023) 100050.
[12] K. P. Yancey, G. Laflair, A. Verardi, J. Burstein, Rating short l2 essays on the cefr scale with gpt-4, in: Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023), 2023, pp. 576–584.
[13] S. Bannò, H. K. Vydana, K. Knill, M. Gales, Can gpt-4 do l2 analytic assessment?, in: Proceedings of the 19th workshop on innovative use of NLP for building educational applications (BEA 2024), 2024, pp. 149–164.
[14] M. V. Cantero-Romero, M. E. Vallecillo-Rodríguez, A. M. Ortiz-Colón, S. M. Jiménez-Zafra, Gpt-3.5 para la nivelación de ele en un entorno universitario: Un estudio comparativo de zero-shot learning y fine-tuning., Procesamiento del Lenguaje Natural 75 (2025).
[15] D. Blanchard, J. Tetreault, D. Higgins, A. Cahill, M. Chodorow, Toefl11: A corpus of non-native english, ETS Research Report Series 2013 (2013) i–15. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.2013.tb02331.x. doi:https://doi.org/10.1002/j.2333-8504.2013.tb02331.x.
[16] Universidad de Santiago de Compostela, Corpus de aprendices de español (CAES), 2022. URL: https://galvan.usc.es, accessed: 2026-05-01.
[17] C. Lozano, Cedel2: Design, compilation and web interface of an online corpus for l2 spanish acquisition research, Second Language Research 38 (2022) 965–983.
[18] A. M. C. Mancera, I. P. Martínez, A. B. Canales, L. C. Fernández, J. F. S. Granda, Corpus para el análisis de errores de aprendices de e/le (corane) (2002) 527–534.
[19] L. Vásquez-Rodríguez, P.-M. Cuenca-Jiménez, S. E. Morales-Esquivel, F. Alva-Manchego, A benchmark for neural readability assessment of texts in spanish, in: Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), 2022.
[20] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, arXiv preprint arXiv:2402.05672 (2024).
[21] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018).
[22] D. Tamayo, I. Lacunza, P. Rivera-Hidalgo, S. Da Dalt, J. Aula-Blasco, A. Gonzalez-Agirre, M. Villegas, Mrbert: Modern multilingual encoders via vocabulary, domain, and dimensional adaptation, arXiv preprint arXiv:2602.21379 (2026).
[23] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, et al., Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 2526–2547.
[24] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, N. Smith, Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, arXiv preprint arXiv:2002.06305 (2020).
[25] M. Mosbach, M. Andriushchenko, D. Klakow, On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines, arXiv preprint arXiv:2006.04884 (2020).

A. Statistics of the Training Corpus

Table 3
Distribution by corpus, task and level

Corpus	Task	A1	A2	B1	B2	C1	TOTAL
CAES	Job change	2425	1	1	0	1	2428
	Letter to a friend	3	1	1926	2	0	1932
	Family	2160	0	0	1	0	2161
	Smoking in public places	0	2	0	1338	0	1340
	Funny story	1	1	1228	0	1	1231
	Note for being late	946	0	0	0	0	946
	Person you admire	0	1691	0	0	0	1691
	Holiday postcard	5	2772	2	0	0	2779
	Complaint to an airline	1	0	1326	0	0	1327
	Complaint to a gas company	2	0	0	0	691	693
	Hotel room reservation	1	1238	0	0	0	1239
	Film review	7	2	0	0	1033	1042
	Admission request	2	0	0	1759	0	1761
CEDEL2	Region where you live	25	153	111	78	58	425
	Famous person	26	128	47	34	18	253
	Film	15	20	35	44	44	158
	Last year holidays	4	36	64	75	88	267
	Future plans	11	48	52	49	48	208
	Recent trip	1	12	25	49	81	168
	An experience	0	4	1	17	19	41
	Terrorism	0	5	4	3	4	16
	Anti-smoking law	3	2	5	7	11	28
	Gay couples	0	0	2	19	32	53
	Marijuana legalization	1	1	1	5	5	13
	Immigration	0	3	1	11	12	27
	Frog	5	48	63	69	97	282
	Chaplin	25	154	350	506	575	1610
CORANE	Type 1	0	38	409	108	322	877
	Type 2	0	5	22	18	35	80
TOTAL		**5669**	**6365**	**5675**	**4192**	**3175**	**25076**

Note: CORANE tasks are classified into Type 1 (tasks completed with supporting material) and Type 2 (tasks completed without supporting material). Within each type, topics are mixed and unlabeled.

Table 4
Mean text length by level (in words)

Level	Mean	Std
A1	105.83	55.58
A2	143.64	65.60
B1	175.90	84.02
B2	244.74	107.30
C1	306.38	168.87