LANG	SENT	PARA	DOC	DIAG
EN	200	200	200	0
ES	0	200	200	0
DE	200	200	200	0
NL	0	0	200	0
CS	0	200	0	0
IT	0	200	0	0
FR	200	0	200	0
ET	0	200	200	0
PT	0	200	0	0
AR	200	200	0	0
HI	200	0	0	0
RU	200	0	0	0
CY	200	109	41	115
Total	**1,400**	**1,709**	**1,241**	**115**

Table 13: Data statistics of **UNIVERSALCEFR-TEST** in terms of levels (sentence, paragraph, document, dialogue) across the 13 target languages.

L1 backgrounds with generally low proficiency. This makes it more difficult for models to identify consistent patterns due to strong L1 interference and low coverage. Overall, both fine-tuned and feature-based models seem to be unable to distinguish between sublevels, with most examples of both A levels being predicted as A1, and the remainder (mostly examples of the B levels) as B1. On the positive side, contrary to what was observed for other languages, the models do not seem to be influenced by text length, with the predictions of XML-R having a correlation of just 0.39 with that feature. The prompting approaches lead to a bias towards the prediction of levels A2 and B1, with the top performer among these approaches (Gemma3 with EN-WRITE prompt) predicting A2 for 28% of the examples and B1 for 62%. Notably, when using the more descriptive prompts, the Gemma 1 model outperformed EuroLLM, in spite of having fewer parameters and not being specifically trained on Portuguese data.

French. The French corpus and our analysis are divided into sentence-level and document-level data. The sentence-level set contains 1,668 sentences ranging from A1 to C2, while the document-level set includes 344 documents from A1 to C1, with an intense concentration at the B levels (75% of the data falls within B1 and B2). In line with the other languages, XLM-R is the most consistent model and achieves the best global performance in every setting. Random Forest (RF) with all features fluctuates more in overall performance, dropping notably in the document-level task, but retains some consistency in terms of which proficiency levels it performs best or worst on. RF with top features performs inconsistently overall but achieves the best results on the document-level task. However, it shows instability in class-level performance, with changes in which levels are most accurately predicted. Among the prompt-based models, Gemma3 is more stable than Gemma1, but both remain below the performance of XLM-R and RF, showing a weaker performance in the LLMs (Gemma1 and Gemma3). Gemma1, in particular, is the least consistent model, with highly variable class-level performance and occasional zero F1 scores for some levels in specific setups. The Gemma1 results are likely due to the lack of French documents during the training of this model. Across all models, prediction is generally more reliable for intermediate levels (A2–B2), while C-level predictions remain the most challenging. Fine-tuning has the clear advantage: the fine-tuned XLM-R achieves the highest accuracy across all evaluation set-ups, making it the most reliable in correctly predicting gold labels. It consistently outperforms all other models, both at the sentence and document levels. This is consistent with previous experiments on French (Yancey et al., 2021; Ngo and Parmentier, 2023; Wilkens et al., 2024), although our performance is slightly lower than in those studies. Prompting is the least effective: both Gemma1 and Gemma3, used in a prompt-based setting, show the lowest prediction accuracy, often failing to identify the correct labels, especially at the extremes of the proficiency scale (A1, C1, and C2 levels). Traditional supervised classifiers (Random Forest) perform moderately well, consistently outperforming the prompt-based models but still lagging behind the fine-tuned model. The feature-based models had a particularly poor performance on C1 and C2 levels. This is likely due to a lack of specialized features for those proficiency levels. Moreover, their performance varies by set-up, with some gains at the document level but noticeable drops elsewhere. Nevertheless, the two RF flavours had similar results. In summary, fine-tuning yields the best predictions, followed by traditional supervised learning, while prompting underperforms in this task.

Model	EN	ES	DE	NL	CS	IT	FR	ET	PT	AR	HI	RU	CY	Tally
GEMMA1	✓													1/1
GEMMA3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	13/140
EUROLLM	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		12/35
MODERNBERT	✓													1/1
EUROBERT	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓		10/15
XLM-R	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	13/100

Table 14: Mapping of language coverage of training data used for the six large, pretrained language models in the model evaluation paradigm in Section 5. Models in **teal** are English-centric (trained primarily with English data), and models in **purple** are multilingual (trained with massive multilingual data). We referred to each model's corresponding release papers and publications for information on their supported languages. Note that the documentation of GEMMA3 indicates it has been trained with 140+ languages. Thus, we loosely consider it to cover all 13 languages in UNIVERSALCEFR. The tally column indicates `{lang_covered/lang_seen}`. For example, EUROBERT covers 10 of the languages in the current UNIVERSALCEFR the 15 languages it supports.

German. For German, the fine-tuned models (>70%) have been shown to outperform all other approaches, such as feature-based (≈50%-65%) and prompting (≈38%-46%), despite the presence of unbalanced CEFR levels in both the training and test data. The findings derived from the English-only and multilingual models, including fine-tuning and prompting methodologies, exhibit no notable difference. This may be due to the similarities between English and German, both of which are West Germanic languages. Alternatively, the great transferability of the fine-tuned English-only model may also be due to the large amount of German training data available (27,000 training samples). The feature-based models performed second best and were still able to compete with the fine-tuned models to some extent. This is surprising, given that a previous analysis showed that the features only exhibited low correlations with CEFR levels (see Section E.3). Proficiency assessment for German appears to require certain idiosyncratic features. For example, the feature covering the maximum distance between words in a dependency tree showed a high feature importance only for German, reflecting the language's free word order and long-distance dependencies. For the prompting setup, the multilingual Gemma3 model performed, achieving good results for lower CEFR levels, but underpredicting higher levels. By contrast, Gemma1 significantly overpredicts level A1 (250 against 14 from the gold labels), resulting in poorer performance on average and across the other levels. One deceptive indicator might be the length of the texts to be classified, as reflected by the strong correlation between text length and Gemma1's predictions (r=0.61). When comparing the prompting setups with regard to language-specific task descriptions, no clear trend emerges across all three LLMs, mirroring the difficulty of prompt engineering for a complex task such as multi-lingual proficiency classification.

Arabic. Across the 400 Arabic test items, Gemma1 tends to over-predict lower CEFR levels, assigning 31 items to A1 while only 12 are from the true labels, and 90 to A2 against 26. There is also a tendency to under-predict C1, with 18 predictions against 40 from the true labels, resulting in the highest average grade deviation of 1.0. In contrast, XLM-R and both Random Forest variants distributed their predictions more evenly overall, with XLM-R achieving the smallest average grade deviation of 0.75. In terms of granularity, the Arabic subset is split into sentence-level, reference data, and paragraph-level learner data. For the sentence-level reference texts, XLM-R (≈55%) and Random Forest models from the two linguistic feature setups (≈49.3%-51.2%) outperform both Gemma1 and Gemma3 models through prompting (≈16.5%-32%). However, with paragraph-level learner texts, Gemma3 leads the evaluation (≈41%). At the same time, XLM-R and the Random Forest models fall behind (≈32%), possibly due to the Arabic data used in the training split, which are entirely sentence-level. In contrast, the Gemma3 model has most likely seen diverse online Arabic data.

D Standardized Dataset Fields

We present the standardized JSON format used as a template when processing all qualified datasets in UNIVERSALCEFR. This structured format ensures flexibility and interoperability into other formats

accepted and used by the AI community, including Huggingface and Croissant. Moreover, this format captures the dimensions that are essential to each instance of CEFR-labeled text, including format or granularity, category, license, and language.

E Full Linguistic Feature Analysis

E.1 All Linguistic Features

Overall, we have extracted **100 diverse linguistic features** which can be grouped into morphosyntactic (62), syntactic (18), length-based (11), lexical (4), readability (2), psycholinguistic (2), and discourse (1). The full list of features, including short descriptions, is available in Appendix E. We extracted a diverse set of 100 linguistic features based on sentence-based linguistic annotation with `spacy` (Montani et al., 2023) and `stanza` (Qi et al., 2020), including tokenization, part-of-speech tagging, and dependency parsing performed. Additionally, we use `fasttext` embeddings (Grave et al., 2018), `pyphen` for hyphenation (Berendsen and Kozea, 2025) and `MEGA.HR` crossling lexicon[9] for imageability and concreteness (Ljubešić, 2018). Most of the features have already been implemented in the `text-simplification-evaluation` (`TSEval`) package[10] (see Martin et al. (2018) for the original version and Stodden and Kallmeyer (2020) for the multilingual version).

In Table 21, we provide an overview of all features including a short description, resources used, and correlation with the CEFR level.

E.2 Top Linguistic Features

To extract the top linguistic features (TOPFEATS), we selected those that are present in the top 10 ranked most important features for at least three languages. Using this criteria, we came up with a list of 23 linguistic features as reported in Table 16 which was then used in the experiment result in Table 4.

E.3 Linguistic Correlation Analysis

In the following, we describe some insights into linguistic diversity of the UniversalCEFR data by correlation analysis between the features and the CEFR levels.

Correlation Across All Languages. Considering the absolute Spearman correlation between the features and the CEFR level (selecting values with $p < 0.05$ and $\rho > 0.3$ on average across all languages), the strongest associations were found in length-based measures, such as characters per sentence and syllables per sentence. Several grammatical complexity features, including parse tree height and phrase length, showed moderate correlations. Readability indices (FKGL and Flesch Reading Ease) also displayed moderate correlations in the expected direction. Psycholinguistic features, such as concreteness and imageability, were negatively correlated with proficiency, indicating a shift toward more abstract language at higher levels. Finally, morphosyntactic features regarding voice, tense, and number showed moderate but consistent correlations, supporting their relevance in reflecting syntactic development.

Correlation By CEFR Level. To assess the consistency of feature relevance across languages, we examined the number of features with significant correlations ($p < 0.05$) with CEFR levels per language. The results revealed notable variations. Languages such as Czech (CS), Estonian (ET), and Italian (IT) showed a high number of relevant features, suggesting strong alignment between the selected linguistic features and CEFR progression in these languages. English (EN), Spanish (ES), French (FR), Hindi (HI), and Russian (RU) showed moderate coverage, with a reasonable number of features exceeding the 0.3 correlation threshold. In contrast, Arabic (AR), Dutch (NL), and Portuguese (PT) exhibited weak coverage, while Welsh (CY) and German (DE) had very few or no features with relevant correlations, indicating a limited match between the current feature set and CEFR levels for those languages. Furthermore, a few features are only relevant for a few languages, e.g., the translative case for only Estonian, negative verb polarity for only Czech, or genitive case for only Czech, Estonian, and Russian. This variability highlights the influence of language-specific properties on the effectiveness of general feature-based models for proficiency prediction.

Point-Biserial Correlation. A point-biserial correlation analysis by CEFR level revealed that most features exhibit only weak correlations, suggesting limited discriminative power when isolating individual CEFR bands. Interestingly, the absolute correlation values tend to be strongest at the

[9] https://www.clarin.si/repository/xmlui/handle/11356/1187

[10] https://github.com/facebookresearch/text-simplification-evaluation