3.1 Design Principles

Our methodology was guided by three key design principles.

Openness and Accessibility. In building UNIVERSALCEFR, we aim to demonstrate how data-driven research in language proficiency and assessment benefits from standardized, unified data formats. This enables portability and interoperability across domains with evolving data pipelines, such as language model pre-training in NLP. All corpora included in UNIVERSALCEFR are publicly available for non-commercial research through permissive licenses (e.g. Creative Commons). However, significant effort was required to collate and standardize these datasets, highlighting the need for standardization and improved accessibility.

Multilinguality and Structure Diversity. Although CEFR originated in Europe, it has been increasingly adopted as a reference framework for language proficiency assessment worldwide. Accordingly, UNIVERSALCEFR extends beyond European languages. Its current version includes 13 languages, spanning high-resource (English, Spanish, French, German, Italian, Portuguese), mid-resource (Dutch, Russian, Arabic), and low-resource (Czech, Estonian, Hindi, Welsh) languages. It also captures structural diversity by annotating each corpus with its production category (learner or reference), granularity (sentence, paragraph, document, or discourse), and label coverage (standard CEFR or CEFR plus levels).

Global Collaboration. From its conceptualization and planning, the UNIVERSALCEFR initiative involved close collaboration among 20 researchers in language proficiency assessment, NLP, and education from 13 institutions across nine countries (UK, Canada, USA, Germany, Sweden, UAE, Spain, Belgium, and Portugal).[7] They all played a key role in defining the standardization protocol, designing evaluation experiments, and discussing future research directions. These collaborative decisions are detailed in the following sections.

3.2 Data Collection

This section outlines the corpus selection criteria and the standardization methods used in UNIVERSALCEFR for acquiring and consolidating a large and diverse collection of resources.

[7] As CEFR is a European framework, most active researchers in the field are based in Europe.

Corpora Selection. The inclusion of datasets in UNIVERSALCEFR is guided by three criteria:

1. **Public Accessibility:** Datasets must be available under a permissive license for non-commercial research (e.g., Creative Commons, CC-BY-NC), or be in the public domain and acquirable through direct download or via a request form for usage tracking.

2. **Gold-Standard CEFR Labels**: Datasets must include CEFR annotations produced or validated by domain experts, such as language teachers or proficiency researchers, particularly in the case of learner texts.

3. **Human Authorship:** All texts must be written by humans to ensure suitability for research involving creative, multilingual, multi-level, and multi-genre content. As of this writing, UNIVERSALCEFR does not include machine-generated texts.

The full list of consolidated corpora that meet all three UNIVERSALCEFR inclusion criteria is provided in Table 22 in the Appendix.

Standardization Process. To ensure interoperability, transformation, and machine readability, we standardized the collected datasets by preprocessing their varied source formats into a unified structure. We adopted JSON as the per-instance format and defined eight metadata fields considered essential for each CEFR-labeled text. These fields include the source dataset, language, granularity (document, paragraph, sentence, discourse), production category (learner or reference), and license. Full descriptions and predetermined values used for each field are provided in Table 15. The final standardized dataset is available from HuggingFace Dataset repository.[8] A key challenge was the lack of a unified format across the language proficiency community. Source corpora came in various formats, including plain text (e.g., csv, tsv, txt), spreadsheets (e.g., XLSX, XLS), markup (e.g., XML), and PDFs requiring manual extraction. This challenge further motivates the need for unified data aggregation initiatives that UNIVERSALCEFR aims to help establish.

[8] https://huggingface.co/UniversalCEFR/datasets

UNIVERSALCEFR	# of Instances
- FULL*	505,807
(out-of-scope instances)	11,316
- FULL	494,491
- TRAIN	435,919
- DEV	54,107
- TEST	4,465

Table 2: Data splits for UNIVERSALCEFR. FULL* denotes all instances, including those with CEFR labels that we currently do not recognize for the task (e.g., NA, A+, B). These were excluded from the TRAIN, DEV, and TEST sets used in our experiments.

3.3 Dataset Statistics

The final UNIVERSALCEFR collection comprises **505,807 CEFR-labeled texts** across **13 languages** and **4 scripts** (Latin, Arabic, Devanagari, and Cyrillic). Tables 2 and 3 show the overall dataset size, its splits and breakdown per CEFR level per language. We identified 11,316 instances with invalid or out-of-scope labels (e.g., NA, A+, B) outside the six recognized CEFR labels (A1–C2) and duplicates, which were removed before splitting UNIVERSALCEFR into TRAIN, DEV, and TEST. For the TEST, we set a cap of 200 instances per language and per granularity level. Additional dataset statistics can be found in Appendix A.

LANG	A1	A2	B1	B2	C1	C2
EN	192,596	132,614	66,425	23,266	8,004	795
ES	8,282	8,648	6,835	5,061	3,224	0
DE	319	15,970	15,630	474	130	426
NL	51	216	782	738	219	85
CS	1	188	165	81	4	0
IT	29	381	394	2	0	0
FR	151	390	575	478	293	126
ET	0	395	588	407	307	0
PT	314	325	367	233	112	72
AR	81	259	625	645	361	183
HI	263	283	286	263	222	174
RU	402	293	409	326	237	91
CY	764	608	0	0	0	0
Total	**203,253**	**160,570**	**93,081**	**31,974**	**13,113**	**1,952**

Table 3: Data statistics of **UNIVERSALCEFR-FULL** in terms of recognized CEFR levels (A1, A2, B1, B2, C1, C2) across the 13 target languages.

4 Linguistic Feature Analysis

We aim to examine how well a broad set of linguistic features aligns with CEFR proficiency levels across languages in UNIVERSALCEFR. We extracted a set of **100 linguistic features**, grouped into morphosyntactic (62), syntactic (18), length-based (11), lexical (4), readability (2), psycholinguistic (2), and discourse (1) categories. A complete and detailed list is available in Appendix E.

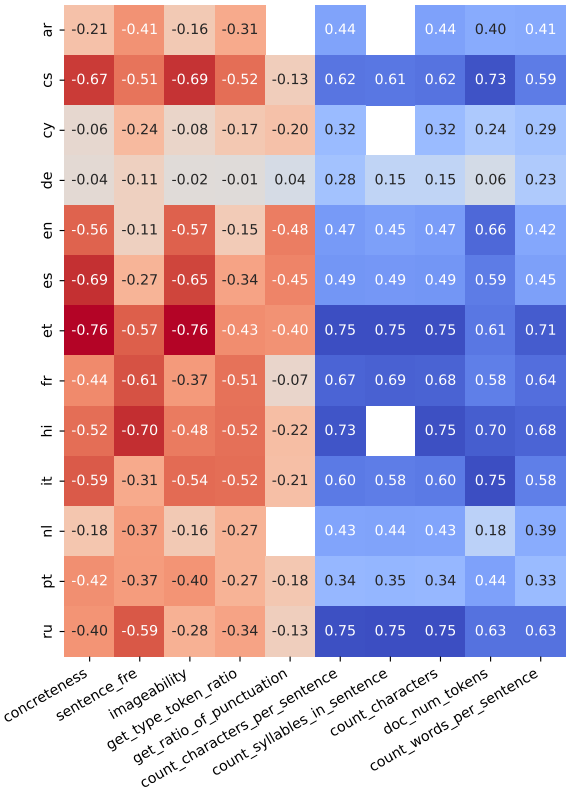

Figure 2: Highly correlated linguistic features occurring in at least three languages. **Blue features** lean towards positive correlation, while **red features** denote negative correlation. For brevity, these top features are those lying at the extreme ends of the correlation spectrum.

4.1 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.

4.2 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 as visualized in Figure 2. 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.

5 CEFR Level Classification

Given the availability of gold-standard CEFR labels and the linguistic diversity of the UNIVERSAL-CEFR dataset, we define our primary experimental task as **multiclass, multilingual CEFR level classification**. The goal is to predict one of the six CEFR levels (A1–C2) for a given text instance in any of the 13 supported languages. We evaluate three modeling paradigms: feature-based classification, fine-tuning of multilingual pre-trained models, and prompting LLMs.

5.1 Feature-Based Models

We evaluated two widely-used classification models from Scikit-Learn (Pedregosa et al., 2011): **Random Forest** (RANDFOREST) and **Logistic Regression** (LOGREGR). Both models were trained on the linguistic features described in Section 4, using Scikit-Learn's default hyperparameter settings. We experimented with two feature configurations: one using all 100 features (ALLFEATS) and another using an automatically selected subset of top-performing features across all languages (TOPFEATS). Appendices E.1 and E.2 detail the linguistic feature information for both setups.

5.2 Fine-tuned Models

We used three BERT-based models with varying degrees of multilingual coverage: **ModernBERT** (Warner et al., 2024), a monolingual English model with 395M parameters; **EuroBERT** (Boizard et al., 2025), a multilingual model trained on 15 diverse European and non-European languages, with 210M parameters; and **XLM-R** (Conneau et al., 2020), a massively multilingual model supporting 100 languages, with 279M parameters. Each model was fine-tuned for three epochs, with the best checkpoint selected based on the highest weighted F1 score on the validation set. Additional details can be found in Appendix Table 17.

5.3 Descriptor-Based Prompting

We evaluated three instruction-tuned models: **Gemma 1** (Gemma Team, 2024), an English-centric model with 7B parameters; **Gemma 3** (Gemma Team, 2025), a multilingual model trained on 140+ global languages with 12B parameters; and **EuroLLM** (Martins et al., 2024), a multilingual model trained on 15 European-centric languages with 9B parameters. We explored five prompting strategies, ranging from no context to setups using CEFR level descriptors for reading comprehension and written production, either in English or in specific languages. The prompt configurations are as follows:

- **BASE**. Generic prompting with no CEFR level descriptors as context.
- **EN-READ**. CEFR level descriptors for reading comprehension in English used as context.
- **EN-WRITE**. CEFR level descriptors for written production in English used as context.
- **LANG-READ**. CEFR level descriptors for reading comprehension, translated to the target language being assessed used as context.
- **LANG-WRITE**. CEFR level descriptors for written production, translated to the target language being assessed used as context.

All CEFR descriptors were retrieved from the official CEFR website. Prompt templates and hyperparameter values for each setup are detailed in Table 18 and Appendix I.

5.4 Evaluation Metrics

We use **weighted F1** as the primary evaluation metric across all experiments. This accounts