# UniversalCEFR: Enabling Open Multilingual Research on Language Proficiency Assessment

**Joseph Marvin Imperial**[1,3], **Abdullah Barayan**[2], **Regina Stodden**[4], **Rodrigo Wilkens**[5], **Ricardo Muñoz Sánchez**[6], **Lingyun Gao**[7], **Melissa Torgbi**[1], **Dawn Knight**[2], **Gail Forey**[1], **Reka R. Jablonkai**[1], **Ekaterina Kochmar**[8], **Robert Reynolds**[9], **Eugénio Ribeiro**[10,11], **Horacio Saggion**[12], **Elena Volodina**[6], **Sowmya Vajjala**[13], **Thomas François**[7], **Fernando Alva-Manchego**[2], **Harish Tayyar Madabushi**[1]

[1]University of Bath, [2]Cardiff University, [3]National University Philippines,
[4]Bielefeld University, [5]University of Exeter, [6]University of Gothenburg, [7]UCLouvain,
[8]MBZUAI, [9]Brigham Young University, [10]INESC-ID Lisboa,
[11]Instituto Universitário de Lisboa (ISCTE – IUL), [12]Universitat Pompeu Fabra,
[13]National Research Council, Canada

Correspondence: jmri20@bath.ac.uk, alvamanchegof@cardiff.ac.uk

## Abstract

We introduce **UniversalCEFR**, a large-scale multilingual multidimensional dataset of texts annotated according to the CEFR (Common European Framework of Reference) scale in 13 languages. To enable open research in both automated readability and language proficiency assessment, UniversalCEFR comprises **505,807 CEFR-labeled texts** curated from educational and learner-oriented resources, standardized into a unified data format to support consistent processing, analysis, and modeling across tasks and languages. To demonstrate its utility, we conduct benchmark experiments using three modelling paradigms: a) linguistic feature-based classification, b) fine-tuning pre-trained LLMs, and c) descriptor-based prompting of instruction-tuned LLMs. Our results further support using linguistic features and fine-tuning pretrained models in multilingual CEFR level assessment. Overall, UniversalCEFR aims to establish best practices in data distribution in language proficiency research by standardising dataset formats and promoting their accessibility to the global research community.

universalcefr.github.io
huggingface.co/UniversalCEFR
github.com/UniversalCEFR

## 1 Introduction

Language proficiency research plays a central role in education, and often intersects with advances in linguistics and artificial intelligence (AI). In natural language processing (NLP), language proficiency has been approached through well-established tasks such as automated readability assessment (ARA) and automated essay scoring (AES). ARA focuses on determining whether a given text matches the expected reading skills of language learners according to their level, whereas AES evaluates the writing skills of the learners as reflected in a text they have written. In this paper, we combine these tasks under the more generic term of *language proficiency assessment*, as it has varied practical applications in educational assessment and calibration of reading materials for learners (Xia et al., 2016; Harsch, 2014; Figueras, 2012) as well as for various NLP tasks (see use cases in Figure 1). A widely recognized standard for measuring second language (L2) proficiency is the Common European Framework of Reference for Languages (CEFR),[1] developed by the Council of Europe. CEFR offers a

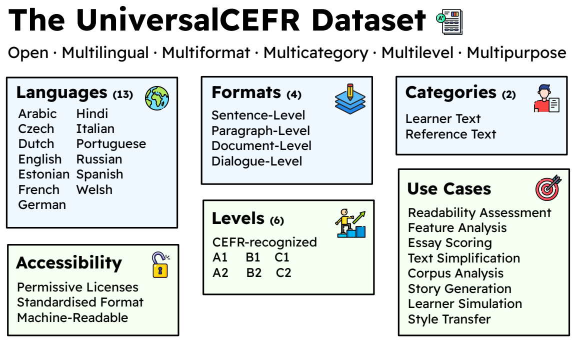

Figure 1: Overview of the contributions of the **UniversalCEFR** dataset, highlighting its **diverse structural coverage**—spanning language, format, category, and CEFR level—as well as its **accessibility and interoperability** for downstream tasks and use cases enabled by permissive licenses and standardized data formats.

[1] https://www.coe.int/en/web/common-european-framework-reference-languages

| Resource | # Datasets Indexed | # Languages Covered | Data Types | Data Accessibility | Standard Format | Geographic Restrictions |
|---|---|---|---|---|---|---|
| CEFRLex | 7$^\dagger$ | 6 | text | unrestricted | no | none |
| Corpora @ UCLouvain | 31$^\dagger$ | 9 | text, audio, video | request per corpus | no | yes |
| CLARIN L2 Learner Corpora | 75$^\dagger$ | 34 | text, video | request per corpus | no | yes |
| Learner Language (Språkbanken) | 15$^\dagger$ | 13 | text, audio | request per corpus | no | yes |
| **UNIVERSALCEFR** | **26** | **13** | **text** | **unrestricted** | **yes** | **none** |

Table 1: Comparison of existing language learning and language proficiency dataset collections with UNIVERSALCEFR. $^\dagger$ indicates that only a small subset of the corresponding resource in that repository contains CEFR labels. Among the five repositories, UNIVERSALCEFR is the only non-geo-locked and standardized collection, allowing seamless, unrestricted use for non-commercial research with proper attribution.

language-independent guide for evaluating learners' abilities in reading, writing, listening, and speaking. It defines a six-level scale (A1, A2, B1, B2, C1, and C2) denoting increasing language competency (North, 2014, 2007).

Recent advances in language proficiency assessment have moved from models relying on handcrafted linguistic features to large language models (LLMs), which achieve high performance across diverse predictive and generative tasks through post-training techniques such as supervised fine-tuning (Devlin et al., 2019; Vaswani et al., 2017) or instruction tuning (Wei et al., 2022). This form of task generalization enables complex linguistic pattern (e.g., features that make a text complex) modelling within unified frameworks for assessing language proficiency on standardized scales like CEFR. Moreover, they can also be extended to low-resourced languages, potentially improving automatic assessment through techniques such as cross-lingual transfer (He and Li, 2024; Imperial and Kochmar, 2023a,b; Vajjala and Rama, 2018).

To fully leverage the potential of modern approaches for CEFR-level prediction, researchers require access to high-quality datasets with broad coverage across languages, proficiency levels, and text granularity. However, despite the long-standing use of CEFR in educational and NLP research, there are very limited standardized, machine-readable, and openly accessible collections of CEFR-annotated corpora, especially in terms of language coverage and granularity beyond sentence level (Naous et al., 2024). Moreover, most existing single-language resources are available in inconsistent or outdated formats (e.g., unprocessed text files, XML), which require extensive preprocessing and normalization. Finally, many datasets are restricted by copyright or licensing terms, limiting their accessibility for open research.

To this end, our work addresses the resource gap in CEFR-based language proficiency assessment research through the following contributions:

- We introduce UNIVERSALCEFR, a large-scale multilingual multidimensional open dataset composed of 505K CEFR-labeled texts across 13 languages, designed to advance multilingual research in language proficiency assessment.
- We propose a data standardization pipeline and annotation template to homogenize available CEFR-labeled texts, enhancing their interoperability and accessibility for researchers across domains.
- We provide a critical reflection of current practices in data sharing of language proficiency assessment resources and suggest pathways towards improvement using UNIVERSALCEFR as a case study for a more open, standardized initiative for resource development.

## 2 Background

**Language Learning Databases and Resources.** Language learning and language proficiency are research areas driven by the collection of two main types of data: reference-based data created by experts (e.g. reference reading materials) and learner-based data created by language learners (e.g. essays, conversations, and dialogue snippets). If a task requires it, such as in proficiency assessment, these corpora may undergo examination by language proficiency experts who will grade them based on a scale (e.g. CEFR). We list four community-recognized databanks and resource collections in the domain of language learning and proficiency assessment in Table 1. CEFRLex is a collection of machine-readable multilingual lexicon-based datasets in 6 European languages.

The Corpora Hub hosted by UCLouvain, the Learner Language from Språkbanken Text (SBX), and the L2 Learning Corpora hosted by the Common Language Resources and Technology Infrastructure (CLARIN) are all large collections of general multilingual and multimodal language learner datasets. Not all corpora in these databases are annotated with CEFR labels, and each corpus is associated with a publication detailing how they were collected and built and their specific purpose in language learning research.

**Access Restrictions and Data Privacy Regulations.** Despite the existence of L2 resource collections as listed in Table 1, researchers cannot freely and openly use all datasets hosted in these repositories. CEFRLex,[2] Corpora @ UCLouvain,[3] CLARIN,[4] and Språkbanken Text[5] are hosted under European universities and institutions which means they are under the jurisdiction of EU Data Privacy Laws, particularly the General Data Protection Regulation (GDPR).[6] Thus, learner texts from these collections, written based on personal interactions and containing Personally Identifiable Information (PII), can only be accessed through special legal coordination with the data maintainers. If access is granted, the licensee may also need to provide a proof of PII anonymization that produces a derivation distinct from the original dataset as done in Jentoft and Samuel (2023) for the ASK Corpus (Tenfjord et al., 2006) containing L2 Norwegian CEFR-labeled texts and the International Corpus of Learner Finnish (ICLFI) (Jantunen et al., 2013) containing L2 Finnish CEFR-labeled texts. Moreover, some datasets such as the SweLL Corpus (Volodina, 2024; Volodina et al., 2019, 2016) from Språkbanken Text, composed of Swedish L2 texts with CEFR levels, are geographically licensed and can only be used by institutions within the EU and EEA region. As such, these datasets remain off-limits to any researcher outside of Europe.

**Automatic CEFR Assessment.** The majority of research on automatic classification (or ranking) of texts based on the CEFR scale tends to focus on single-language model evaluations (Ribeiro et al., 2024a; Wilkens et al., 2024, 2023, 2018; Tack et al., 2017; Volodina et al., 2016; Pilán et al., 2016; Vajjala and Lõo, 2014; Xia et al., 2016; Yancey et al., 2021; Vásquez-Rodríguez et al., 2022). This allows deeper investigation of language-specific nuances and intricacies connected to measuring text complexity. Meanwhile, other works have explored universal, language-agnostic features such as Azpiazu and Pera (2019); Arhiliuc et al. (2020); Caines and Buttery (2020); Vajjala and Rama (2018) where they used traditional word and PoS-ngram features to build a multi- and cross-lingual CEFR proficiency classifier for German, Czech, Italian, Spanish, and English, among others. He and Li (2024), on the other hand, focused on cross-lingual automatic essay scoring anchored on the CEFR scale, covering six languages (Czech, English, German, Italian, Portuguese, and Spanish).

In parallel with the rise of benchmarking studies for LLMs, similar efforts are growing in the CEFR-based language proficiency community. Two works in this direction include Naous et al. (2024), which introduced ReadMe++, a multilingual, multidomain dataset for sentence-level readability assessment on a CEFR scale covering five languages, while the iRead4Skills Project by Pintard et al. (2024) released a collection of written texts in French, Portuguese, and Spanish across multiple genres and levels patterned to CEFR.

## 3 The UNIVERSALCEFR Dataset

To support multilingual language proficiency research, we introduce UNIVERSALCEFR, a large-scale initiative that curates and standardizes open human-annotated CEFR-labeled corpora. In this section, we outline the design principles behind dataset's development, the data collection and standardization pipeline, key dataset statistics, and a linguistic feature analysis that supports downstream modeling.

### 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

[2] https://cental.uclouvain.be/cefrlex/

[3] https://corpora.uclouvain.be/catalog/

[4] https://www.clarin.eu/resource-families/L2-corpora

[5] https://spraakbanken.gu.se/en/resources/learner-language

[6] https://gdpr-info.eu/

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 discussed in more detail 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.

**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, multilevel, 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 was uploaded to a HuggingFace Dataset repository,[8] and will be made publicly available after the review period. 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.

### 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). Table 2 shows the overall dataset size and its splits.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

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

[8] https://github.com/huggingface/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.

the TEST, we set a cap of 200 instances per language and per granularity level. Additional dataset statistics can be found in Appendix A.

### 3.4 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.

To explore the relationship between CEFR levels and linguistic variation in UNIVERSALCEFR, we conducted a Spearman correlation analysis. Length-based measures, such as characters and syllables per sentence, showed the strongest correlation with CEFR levels across all languages. However, this varied significantly across languages: Czech, Estonian, and Italian showed a high number of correlated features, whereas Welsh and German showed very few or none, indicating a limited alignment between the current feature set and CEFR levels in those languages.

In an additional point-biserial correlation analysis by CEFR level (i.e., one level versus all others), most features showed only weak correlations, suggesting limited discriminative power when isolating individual CEFR bands. However, we found an interesting dynamic in some features, e.g., the word length in characters was negatively correlated with A1, neutral with A2, and positive with B1 and higher levels. Hence, the directionality of several features suggests dynamic usage patterns across CEFR bands, even if the correlation strengths remain modest. Further results and detailed analyses are provided in Appendix E.3.

## 4 CEFR Level Classification

Given the availability of gold-standard CEFR labels and the linguistic diversity of the UNIVERSALCEFR 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.

### 4.1 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 3.4, 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). Appendix E.1 and E.2 detail the linguistic feature information for both setups.

### 4.2 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.

### 4.3 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

| Model & Setup | EN | ES | DE | NL | CS | IT | FR | ET | PT | AR | HI | RU | CY | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Baseline** | | | | | | | | | | | | | | |
| Most Frequent Class | 7.39 | 18.1 | 26.8 | 21.4 | 23.8 | 35.5 | 16.3 | 15.9 | 10.0 | 23.3 | 7.28 | 10.7 | 33.4 | 19.3 |
| **Gemma1-7B (English)** | | | | | | | | | | | | | | |
| Base | 21.8 | 26.0 | 40.6 | 32.1 | 44.0 | 57.3 | 32.2 | 39.0 | 14.0 | 28.9 | 25.0 | 34.8 | 48.7 | 34.2 |
| En-Read | 20.5 | 28.3 | 31.0 | 23.5 | 53.6 | 41.0 | 22.7 | 24.9 | 27.2 | 29.5 | 8.4 | 18.0 | 55.7 | 29.6 |
| En-Write | 19.8 | 24.5 | 34.5 | 29.3 | 51.9 | 57.7 | 27.7 | 42.7 | 22.2 | 20.8 | 14.0 | 27.6 | 52.1 | 32.9 |
| Lang-Read | 20.5 | 29.3 | 35.1 | 37.8 | 55.3 | 48.0 | 27.1 | 44.6 | 20.2 | 32.2 | 12.8 | 26.2 | 52.8 | 34.0 |
| Lang-Write | 19.8 | 29.8 | 32.6 | 34.0 | 49.9 | 61.7 | 26.3 | 46.3 | 21.2 | 36.7 | 12.7 | 26.9 | 53.6 | **34.7** |
| **Gemma3-12B (Multi)** | | | | | | | | | | | | | | |
| Base | 28.8 | 35.0 | 42.2 | 47.0 | 42.6 | 65.2 | 38.1 | 39.5 | 24.6 | 41.8 | 28.7 | 29.7 | 40.9 | 38.8 |
| En-Read | 19.3 | 25.5 | 35.8 | 25.5 | 18.5 | 22.9 | 29.3 | 26.0 | 9.8 | 33.3 | 14.8 | 21.2 | 20.5 | 23.3 |
| En-Write | 26.6 | 36.7 | 46.4 | 46.7 | 50.1 | 77.4 | 40.5 | 43.8 | 27.3 | 48.6 | 24.0 | 37.4 | 52.4 | **43.2** |
| Lang-Read | 19.3 | 28.1 | 35.2 | 37.6 | 50.9 | 64.8 | 35.0 | 30.4 | 26.1 | 29.5 | 20.5 | 32.5 | 61.6 | 36.3 |
| Lang-Write | 26.6 | 33.2 | 38.3 | 39.6 | 55.0 | 76.4 | 37.7 | 42.4 | 25.4 | 38.0 | 24.6 | 31.5 | 53.7 | 40.2 |
| **EuroLLM-9B (Multi)** | | | | | | | | | | | | | | |
| Base | 18.6 | 25.4 | 28.0 | 29.1 | 25.0 | 39.9 | 25.9 | 32.0 | 16.4 | 34.3 | 12.7 | 15.1 | 14.4 | 24.4 |
| En-Read | 23.1 | 26.9 | 38.1 | 30.2 | 33.3 | 41.9 | 24.5 | 33.6 | 19.9 | 33.8 | 18.0 | 21.8 | 26.4 | **28.6** |
| En-Write | 21.5 | 26.2 | 29.8 | 32.0 | 32.4 | 33.1 | 26.8 | 32.8 | 21.1 | 31.8 | 17.7 | 17.5 | 24.5 | 26.7 |
| Lang-Read | 23.1 | 27.0 | 32.7 | 31.8 | 29.8 | 32.9 | 28.3 | 28.6 | 16.8 | 32.4 | 14.3 | 16.2 | 17.3 | 25.5 |
| Lang-Write | 21.5 | 28.5 | 35.1 | 30.1 | 30.8 | 30.6 | 27.6 | 29.9 | 16.5 | 35.2 | 21.0 | 16.1 | 8.80 | 25.5 |
| **fine-tuned Models** | | | | | | | | | | | | | | |
| ModernBERT (English) | 75.8 | 71.8 | 72.1 | 54.2 | 66.9 | 82.7 | 47.2 | 88.3 | 33.5 | 30.8 | 51.6 | 48.9 | 73.2 | 61.3 |
| EuroBERT (Multi) | 74.6 | 72.0 | 70.6 | 53.2 | 63.9 | 79.7 | 42.0 | 86.6 | 32.1 | 35.4 | 44.7 | 45.9 | 79.9 | 60.0 |
| XLM-R (Multi) | 75.5 | 69.6 | 73.2 | 59.0 | 68.8 | 83.2 | 51.6 | 88.8 | 29.2 | 43.0 | 52.8 | 49.6 | 72.6 | **62.8** |
| **Feature-Based Models** | | | | | | | | | | | | | | |
| RandForest (TopFeats) | 62.0 | 57.6 | 64.9 | 54.5 | 69.5 | 79.9 | 44.1 | 84.2 | 27.8 | 43.8 | 44.1 | 47.2 | 72.9 | 57.9 |
| RandForest (AllFeats) | 63.4 | 60.6 | 65.4 | 53.0 | 69.2 | 79.3 | 41.4 | 84.2 | 26.4 | 42.8 | 46.8 | 47.8 | 78.2 | **58.3** |
| LogRegr (AllFeats) | 32.1 | 28.2 | 50.9 | 47.1 | 62.9 | 81.9 | 41.7 | 67.5 | 23.1 | 34.1 | 47.8 | 41.1 | 63.8 | 47.9 |
| LogRegr (TopFeats) | 30.4 | 29.7 | 52.5 | 44.1 | 62.7 | 82.7 | 40.3 | 67.5 | 22.7 | 33.5 | 48.4 | 41.1 | 59.2 | 47.3 |

Table 3: Full weighted F1 performance results from the multilingual and English-centric model evaluation experiments using three setups (feature-based, fine-tuning, and prompting) and using UniversalCEFR-test split across the 13 languages. **Boldfaced** values indicate the highest scores overall per model setup, while underlined values highlight the highest scores for each model setup within each language.

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 the Table 18 and Appendix I.

### 4.4 Evaluation Metrics

We use **weighted F1 score** as primary evaluation metric across all experiments. This accounts for the class imbalance in CEFR level distribution and granularity across language subsets in UniversalCEFR-test. Using accuracy in the experiments would produce misleading performance in favor of any majority class.

## 5 Results

### 5.1 Model-Based Performance Comparison

Table 3 shows that, in terms of overall average performance across languages, the fine-tuned

| Model | Sent | Para | Doc | All |
|---|---|---|---|---|
| Gemma1 | 19.41 | **42.74** | 30.81 | 33.63 |
| Gemma3 | 38.71 | **43.12** | 39.62 | 42.33 |
| XLM-R | 62.67 | 66.38 | **71.12** | 65.92 |
| RandForest-All | 56.88 | 62.77 | **64.58** | 61.38 |
| RandForest-Top | 53.89 | 62.98 | **64.94** | 60.50 |

Table 4: Weighted F1 scores for top-performing unique model evaluation setups across granularities available for all languages.

setup with ModernBERT, EuroBERT, and XLM-R achieved the highest weighted F1 score range (≈60%-62.8%) outperforming feature-based models (≈47%-58%) and prompting (≈23%-43%). Among the LLM-based approaches—prompting and fine-tuning—models trained on broader multilingual corpora generally performed better. For instance, XLM-R, which supports 100 languages, was the top performer, followed by EuroBERT (15 languages) and ModernBERT (English-only). A similar trend was observed in prompting: Gemma 3, trained on 140+ languages, outperformed EuroLLM (15 languages) and the English-centric Gemma 1, achieving the best prompting score of 43.2. These findings are consistent with previous work (Naous et al., 2024; Shardlow et al., 2024; Colla et al., 2023; Yuan and Strohmaier, 2021), reinforcing the usefulness of multilingual models for language proficiency assessment tasks. One limitation of our experimental setup, however, is that we did not include language-specific pre-trained models for languages other than English, which may have further improved performance for low- and mid-resource languages.

### 5.2 Granularity-Level Comparison

Table 4 highlights clear performance differences across text granularities (sentence, paragraph, and document) for all models, but more prominently for the Gemma models under prompting. Gemma 1, in particular, tends to over-predict lower CEFR levels (A1-B1) on sentence-level data, whereas its predictions on document-level subsets are more evenly distributed and better aligned with ground truth distributions. This suggests that prompt-based methods may require longer texts to make more accurate predictions, unlike models trained or fine-tuned on the respective datasets. For other models, such as XLM-R, Random Forest, performance show better results with document (≈64%-71%) and paragraph-level data (≈62%-66%) than sentence-level data (≈53%-62%), which was shown as a more difficult task in previous work on readability (Dell'Orletta et al., 2011; Vajjala and Meurers, 2014). Regarding language-specific differences, among English, German, and Welsh, the best performance is seen with the paragraph level dataset for English, the document level dataset for German, and the sentence level dataset for Welsh and French with the fine-tuned XLM-R model. Similar variations can be observed for other languages with more than one level of granularity (see Table 19). No single granularity or model gives a consistently better performance across all tested languages. These results are likely due to the distribution of excerpts across granularity levels in each language (see Table 7 in Appendix A).

| Language | Learner | Reference |
|---|---|---|
| AR | 41.92[†] | 54.69 |
| DE | 71.14 | **74.39** |
| EN | 83.41 | 58.24 |
| ES | **97.99** | 42.72 |

Table 5: Average performances of the best models on learner text versus reference text across languages.[†] indicates performance with Gemma3, and the rest are the performance of the XLM-R model. Only these four languages have both learner and reference texts.

### 5.3 Learner-Reference Comparison

Four languages in UniversalCEFR contain both learner and reference texts: Arabic, German, English, and Spanish. Table 5 reports the average weighted F1 performance difference between the two categories across the four languages. For German, performance is comparable between learner and reference texts (≈71–74%). In contrast, English and Spanish show higher performance on learner texts (83% and 98%) than on reference texts (58% and 42%, respectively). Arabic displays the opposite trend: results on reference texts (54%) are much higher than those of learner texts, where the best results was obtained by Gemma 3 (41%). One possible explanation is that Gemma 3 may have been exposed to more Arabic content in its pre- and post-training phases.

## 6 Discussion

We discuss potential pathways through which UniversalCEFR can serve as a model, and offer key considerations for advancing data accessibility in language proficiency research.

**Critical Reflections of Current Practices.** The multiregional and multidisciplinary effort to curate UNIVERSALCEFR provided us with a comprehensive view of current inconsistencies and critical gaps in building CEFR-labeled language proficiency assessment corpora. On examining the annotation practices, there appears to be no standard method for conducting expert annotations, including inconsistent use of inter-annotator agreement metrics and unclear guidelines on the number of annotators required to achieve reliable agreement. This is reflected in the UNIVERSALCEFR dataset, where nearly half of the corpora lack information on the annotators involved and their agreement scores. We posit that this may be due to diverse judgment of what constitutes high-quality data that does not require further human annotations.

In terms of language coverage, UNIVERSALCEFR includes nine (EN, ES, DE, NL, CS, IT, FR, ET, PT) of the 24 recognized European languages.[9] As a result, researchers working on these nine languages now have access to open, standardized data for CEFR-based language proficiency assessment. The remaining 15 languages represent valuable opportunities for future expansion through collaborative efforts. While our open data and standardization initiative is a step towards addressing current challenges in interoperability and accessibility of resources, similar parallel efforts are needed in areas such as annotation and evaluation practices to ensure sustained progress in the language proficiency assessment community.

**Need for Pro-Research Data Sharing Policies.** As generative AI, particularly LLMs, becomes more ubiquitous, organizations that create valuable data for language proficiency assessment, such as publishers, educational institutions, and media outlets, are growing more cautious about how their resources are used. A major concern is the risk of data being used to train proprietary generative models, especially when such models are only accessible via commercial APIs that require transferring evaluation corpora to external servers. An example is the TCFLE-8 corpus (Wilkens et al., 2023) containing CEFR-labeled essays hosted by France Education International. Researchers seeking access to this dataset must explicitly specify that the resource will not be processed through commercial APIs to prevent potential data harvesting. To address these concerns, we believe the community needs to agree on a unified pro-research data sharing policy with clear usage guidelines for academic, non-commercial studies that require analysis of protected data with generative AI models without training on them.

**Linguistic Features and Fine-tuning Still Matter.** While recent advances in LLMs keep transforming NLP research, our multidimensional experiments in Section 5 reaffirm the continued value of linguistic features for traditional ML classifiers and fine-tuning pretrained models in language proficiency assessment. We observe common patterns where higher distribution and instance count lead to better results using these two setups (see performances on Spanish, English, and German subsets in Table 3) over prompting with CEFR descriptors. Moreover, using linguistic features in language proficiency assessment allows deeper analysis of language interactions with variables such as complexity. For a more detailed discussion, Appendix C provides language-specific analyses across model predictions and possible factors influencing performance, including potential effects of scale, and language exposure from multilingual models.

## 7 Conclusion and Future Directions

In this work, we introduced UNIVERSALCEFR, a large-scale, open, multilingual, multidimensional dataset comprising 505,807 CEFR-annotated texts across 13 languages developed through global collaboration. Our findings through diverse model experiments with CEFR level prediction strengthen support for the utility of linguistic features and fine-tuning multilingual models in language proficiency assessment. Similarly, our critical analysis of the current data and resource-building practices emphasized the need for similar initiatives from the community, and pro-research data sharing policies in the advent of generative AI to remove barriers to accessibility without compromising data privacy and intellectual property.

Beyond its data and technical contributions, UNIVERSALCEFR also carries broader sociolinguistic significance. UNIVERSALCEFR addresses the growing linguistic inequality in modern AI development through focusing on under-represented languages alongside English. We hope this initiative can lead to more responsible AI development that actively resists the growing linguistic centralization around English in global AI research—a mod-

[9] https://european-union.europa.eu/principles-countries-history/languages_en

ern *Matthew effect* (Merton, 1988)—where well-resourced languages receive disproportionate technological attention while smaller languages (like Czech or Welsh) are left behind (Masciolini et al., 2025). The UNIVERSALCEFR is a strong step towards mitigating the Matthew effect.

## Limitations

We discuss several limitations of our work for UNIVERSALCEFR and how researchers can consider these directions to develop the resource further.

**Language Availability and Dependency.** Due to the nature of UNIVERSALCEFR being a collection of open-sourced, publicly accessible CEFR data, its growth depends heavily on how the community will move forward and continuously release artifacts, including CEFR-annotated corpora for reproducibility and wider access for research purposes. We also give credit to the efforts of researchers who work on multi-framework adoption, where CEFR descriptors and bands are overlapped with languages not within Europe (such as Hindi (Naous et al., 2024) and Arabic (Habash and Palfreyman, 2022)), and continue to open-source the annotated data.

**Natural Data Disparity Across Languages.** From the statistics presented in Tables 6 and 7 for UNIVERSALCEFR, it is expected that not all languages have the exact same distribution of data across dimensions. Research efforts on CEFR with the English language remain the portion with the largest open-sourced dataset, being an international language. Moreover, languages such as Welsh do not have any openly accessible CEFR-labeled data at all, to which UNIVERSALCEFR is set to be the first contributor for A1 and A2, covering sentence-, paragraph-, document-, and dialogue-level formats.

**Modalities Beyond Texts.** The current data collection scope of UNIVERSALCEFR and the insights presented in this work only cover CEFR-based texts for now, specifically for reading and writing specifications. Multimodal data, such as audio and video recordings of learners associated with CEFR specifications for listening and speaking, are not yet covered. Naturally, these datasets are even more challenging to acquire and open-source, especially if they contain materials from or are created by learners under legal age and if they contain personal information.

## Ethics Statement

As mentioned throughout this paper, all the datasets we collected for UNIVERSALCEFR based on our criteria presented in Section 3 are already publicly accessible with permissive licenses, and can be used for non-commercial research purposes. While there are three corpora from UNIVERSALCEFR—namely APA-LHA, DEplain, EFCAMDAT—that require users to fill a short form and agree to terms, we still classified them as publicly accessible due to the quick response to access approval.

In the context of the EU AI Act, the use of AI systems for educational purposes is classified under *high risk*, especially those that are intended to *"to evaluate learning outcomes, including when those outcomes are used to steer the learning process of natural persons in educational and vocational training institutions at all levels."* (European Parliament and Council, 2024). Thus, AI systems that will be released in the market with these goals are required to comply with obligations for high-risk systems, including data governance with high-quality, representative datasets. As a form of contribution towards meeting these requirements, the UNIVERSALCEFR is an initiative that will allow researchers and developers access to diverse, multilingual, multidimensional CEFR-labeled texts which can be used for designing systems that are representative, explainable, and fair.

## Acknowledgments

JMI is supported by the National University Philippines and the UKRI Centre for Doctoral Training in Accountable, Responsible, and Transparent AI [EP/S023437/1] of the University of Bath.

## References

Kais ALLKIVI, Pille ESLON, Taavi KAMARIK, Karina KERT, Jaagup KIPPAR, Harli KODASMA, Silvia MAINE, and Kaisa NORAK. 2024. ELLE-Estonian Language Learning and Analysis Environment. *Baltic Journal of Modern Computing*, 12(4).

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cristina Arhiliuc, Jelena Mitrović, and Michael Granitzer. 2020. Language proficiency scoring. In *Proceedings of the Twelfth Language Resources and*

*Evaluation Conference*, pages 5624–5630, Marseille, France. European Language Resources Association.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Wilbert Berendsen and Kozea. 2025. Pyphen. `https://github.com/Kozea/Pyphen`.

Olga Blinova and Nikita Tarasov. 2022. A hybrid model of complexity estimation: Evidence from russian legal texts. *Frontiers in Artificial Intelligence*, 5:1008530.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno M. Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. Eurobert: Scaling multilingual encoders for european languages. *Preprint*, arXiv:2503.05500.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mark Breuker. 2022. Cefr labelling and assessment services. In *European Language Grid: A Language Technology Platform for Multilingual Europe*, pages 277–282. Springer International Publishing Cham.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Andrew Caines and Paula Buttery. 2020. REPROLANG 2020: Automatic proficiency scoring of Czech, English, German, Italian, and Spanish learner essays. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5614–5623, Marseille, France. European Language Resources Association.

Davide Colla, Matteo Delsanto, and Elisa Di Nuovo. 2023. EliCoDe at MultiGED2023: fine-tuning XLM-RoBERTa for multilingual grammatical error detection. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 24–34, Tórshavn, Faroe Islands. LiU Electronic Press.

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.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ–IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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.

European Parliament and Council. 2016. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). `https://eur-lex.europa.eu/eli/reg/2016/679/oj`. OJ L 119, 4.5.2016, p. 1–88.

European Parliament and Council. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts. `https://eur-lex.europa.eu/eli/reg/2024/1689/oj`. OJ L 2024/1689, 12.7.2024, p. 1–88.

Neus Figueras. 2012. The impact of the CEFR. *ELT journal*, 66(4):477–485.

Jeroen Geertzen, Theodora Alexopoulou, Anna Korhonen, and 1 others. 2013. 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*, pages 240–254.

Gemma Team. 2024. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295*.

Gemma Team. 2025. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, Magali Paquot, and 1 others. 2009. *International corpus of learner English*, volume 2. UCL, Presses Univ. de Louvain.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nizar Habash and David Palfreyman. 2022. ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88, Marseille, France. European Language Resources Association.

Claudia Harsch. 2014. General Language Proficiency Revisited: Current and Future Issues. *Language Assessment Quarterly*, 11(2):152–169.

Junyi He and Xia Li. 2024. Zero-shot cross-lingual automated essay scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17819–17832, Torino, Italia. ELRA and ICCL.

Yan Huang, Jeroen Geertzen, Rachel Baker, Anna Korhonen, Theodora Alexopoulou, and EF Education First. 2017. The EF Cambridge open language database (EFCAMDAT): Information for users.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023a. Automatic readability assessment for closely related languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5371–5386, Toronto, Canada. Association for Computational Linguistics.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023b. BasahaCorpus: An expanded linguistic resource for readability assessment in Central Philippine languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6302–6309, Singapore. Association for Computational Linguistics.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2024. SpeciaLex: A benchmark for in-context specialized lexicon learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 930–965, Miami, Florida, USA. Association for Computational Linguistics.

O Jantunen, Sisko Brunni, and University of Oulu, Department of Finnish Language. 2013. International Corpus of Learner Finnish.

Matias Jentoft and David Samuel. 2023. NoCoLA: The Norwegian corpus of linguistic acceptability. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 610–617, Tórshavn, Faroe Islands. University of Tartu Library.

Nikola Ljubešić. 2018. Concreteness and imageability lexicon MEGA.HR-crossling. Slovenian language resource repository CLARIN.SI.

Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Éric de La Clergerie, Antoine Bordes, and Benoît Sagot. 2018. Reference-less quality estimation of text simplification systems. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 29–38, Tilburg, the Netherlands. Association for Computational Linguistics.

Cristina Martins, T Ferreira, M Sitoe, C Abrantes, M Janssen, A Fernandes, A Silva, I Lopes, I Pereira, and J Santos. 2019. Corpus de produções escritas de aprendentes de PL2 (PEAPL2): Subcorpus Português língua estrangeira. *Coimbra: CELGA-ILTEC*.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, and 1 others. 2024. EuroLLM: Multilingual Language Models for Europe. *arXiv preprint arXiv:2409.16235*.

Arianna Masciolini, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfalı, Ricardo Muñoz Sánchez, Elena Volodina, Robert Östling, Kais Allkivi, Špela Arhar Holdt, Ilze Auzina, Roberts Darģis, Elena Drakonaki, Jennifer-Carmen Frey, Isidora Glišić, Pinelopi Kikilintza, Lionel Nicolas, Mariana Romanyshyn, Alexandr Rosen, and 11 others. 2025. Towards better language representation in Natural Language Processing: A multilingual dataset for text-level Grammatical Error Correction. *International Journal of Learner Corpus Research*.

Amália Mendes, Sandra Antunes, Maarten Janssen, and Anabela Gonçalves. 2016. The COPLE2 corpus: a learner corpus for Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3207–3214, Portorož, Slovenia. European Language Resources Association (ELRA).

Robert K Merton. 1988. The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property. *Isis*, 79(4):606–623.

Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spacy: v3.7.2: Fixes for apis and requirements. Version v3.7.2.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.

Duy Van Ngo and Yannick Parmentier. 2023. Towards sentence-level text readability assessment for French. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 78–84, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Brian North. 2007. The CEFR Illustrative Descriptor Scales. *The Modern Language Journal*, 91(4):656–659.

Brian North. 2014. *The CEFR in Practice*, volume 4. Cambridge University Press.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Mathieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2016. A Readable Read: Automatic Assessment of Language Learning Materials based on Linguistic Complexity. *International Journal of Computational Linguistics and Applications (IJLCA)*, 7(1):143–159.

Alice Pintard, Thomas François, Justine Nagant de Deuxchaisnes, Sílvia Barbosa, Maria Leonor Reis, Michell Moutinho, Ricardo Monteiro, Raquel Amaro, Susana Correia, Sandra Rodríguez Rey, Marcos Garcia González, Keran Mu, and Xavier Blanco Escoda. 2024. iRead4Skills Dataset 1: corpora by complexity level for FR, PT and SP (2.1.).

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Robert Reynolds. 2016. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 289–300, San Diego, CA. Association for Computational Linguistics.

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2024a. Automatic text readability assessment in European Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 97–107, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2024b. Avaliação Automática do Nível de Complexidade de Textos em Português Europeu. *Linguamática*, 16(2):121–145.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, and 3 others. 2024. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Itamar Shatz. 2020. Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale English learner language database. *International Journal of Learner Corpus Research*, 6(2):220–236.

Marina Solnyshkina, Vladimir Ivanov, and Valery Solovyev. 2018. Readability formula for russian texts: A modified version. In *Advances in Computational Intelligence: 17th Mexican International Conference on Artificial Intelligence, MICAI 2018, Guadalajara, Mexico, October 22–27, 2018, Proceedings, Part II 17*, pages 132–145. Springer.

Nicolas Spring, Annette Rios, and Sarah Ebling. 2021. Exploring German multi-level text simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1339–1349, Held Online. INCOMA Ltd.

Regina Stodden and Laura Kallmeyer. 2020. A multi-lingual and cross-domain analysis of features for text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 77–84, Marseille, France. European Language Resources Association.

Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16441–16463, Toronto, Canada. Association for Computational Linguistics.

Anaïs Tack, Thomas François, Sophie Roekhaut, and Cédrick Fairon. 2017. Human and automated CEFR-based grading of short answers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179, Copenhagen, Denmark. Association for Computational Linguistics.

Kari Tenfjord, Paul Meurer, and Knut Hofland. 2006. The ASK corpus - a language learner corpus of Norwegian as a second language. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Peter Thwaites, Nathan Vandeweerd, and Magali Paquot. 2024. Crowdsourced Comparative Judgement for Evaluating Learner Texts: How Reliable are Judges Recruited from an Online Crowdsourcing Platform? *Applied Linguistics*.

Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR Level Prediction for Estonian Learner Text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 113–127, Uppsala, Sweden. LiU Electronic Press.

Sowmya Vajjala and Detmar Meurers. 2014. Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 288–297, Gothenburg, Sweden. Association for Computational Linguistics.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

Laura Vásquez-Rodríguez, Pedro-Manuel Cuenca-Jiménez, Sergio Morales-Esquivel, and Fernando Alva-Manchego. 2022. A benchmark for neural readability assessment of texts in Spanish. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 188–198, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30.

Elena Volodina. 2024. On two SweLL learner corpora–SweLL-pilot and SweLL-gold. In *Huminfra Conference*, pages 83–94.

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and 1 others. 2019. The SweLL language learner corpus: From design to annotation. *Northern European Journal of Language Technology (NEJLT)*, 6:67–104.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016. SweLL on the rise: Swedish learner language corpus for European reference level studies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 206–212, Portorož, Slovenia. European Language Resources Association (ELRA).

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

Rodrigo Wilkens, Alice Pintard, David Alfter, Vincent Folny, and Thomas François. 2023. TCFLE-8: a corpus of learner written productions for French as a foreign language and its application to automated essay scoring. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3465, Singapore. Association for Computational Linguistics.

Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, and Thomas François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2316–2331, St. Julian's, Malta. Association for Computational Linguistics.

Rodrigo Wilkens, Leonardo Zilio, and Cédrick Fairon. 2018. SW4ALL: a CEFR classified and aligned corpus for language learning. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Kevin Yancey, Alice Pintard, and Thomas Francois. 2021. Investigating readability of french as a foreign language with deep learning and cognitive and pedagogical features. *Lingue e linguaggio*, 20(2):229–258.

Helen Yannakoudakis, Øistein E Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3):251–267.

Zheng Yuan and David Strohmaier. 2021. Cambridge at SemEval-2021 task 2: Neural WiC-model with data augmentation and exploration of representation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 730–737, Online. Association for Computational Linguistics.

Xuanming Zhang, Zixun Chen, and Zhou Yu. 2024. ProLex: A benchmark for language proficiency-oriented lexical substitution. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8475–8493, Bangkok, Thailand. Association for Computational Linguistics.

## A Full Data Statistics

Tables 7, 9, 11and 13 report the quantity of CEFR-labeled texts across granularity levels per language, and Tables 6, 8, 10 and 12 reflect their counterpart in terms of CEFR level coverage. In forming the TEST split, we randomly sampled CEFR-labeled text instances per language per granularity level, while setting a cap of 200. This allows us to have a sizeable representation of UNIVERSALCEFR while maintaining efficiency for inference with LLMs. In total, we have 4,465 CEFR-labeled instances for UNIVERSALCEFR-TEST, which is comparable to the general sizes of benchmark test sets from previous works related to language proficiency (Naous et al., 2024; Zhang et al., 2024; Imperial and Tayyar Madabushi, 2024). For the TRAIN and DEV sets for fine-tuning and feature-based classification, we split the FULL subset (minus the TEST set) into a 90%-10% partition, respectively.

| 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 6: Data statistics of **UNIVERSALCEFR-FULL** in terms of recognized CEFR levels (A1, A2, B1, B2, C1, C2) across the 13 target languages.

## B Coverage of Large Language Models

In Table 14, we map each model's language coverage or language support based on its respective release papers and publications. Language support means what specific languages have been added and in substantial quantities in a model's training data (e.g., multilingual Wikipedia data dumps for pretraining XLM-R (Conneau et al., 2020)).

| LANG | SENT | PARA | DOC | DIAG |
|---|---|---|---|---|
| EN | 12,826 | 409,362 | 1,837 | 0 |
| ES | 0 | 713 | 31,355 | 0 |
| DE | 26,244 | 1,033 | 5,673 | 0 |
| NL | 0 | 0 | 3,596 | 0 |
| CS | 0 | 441 | 0 | 0 |
| IT | 0 | 813 | 0 | 0 |
| FR | 1,669 | 0 | 344 | 0 |
| ET | 0 | 420 | 1,277 | 0 |
| PT | 0 | 1,423 | 0 | 0 |
| AR | 1,945 | 215 | 0 | 0 |
| HI | 1,491 | 0 | 0 | 0 |
| RU | 1,758 | 0 | 0 | 0 |
| CY | 1,107 | 109 | 41 | 115 |
| **Total** | **47,040** | **414,529** | **115** | **44,123** |

Table 7: Data statistics of **UNIVERSALCEFR-FULL** in terms of levels (sentence, paragraph, document, dialogue) across the 13 target languages.

## C Language-Specific Analysis

We provide in-depth analysis of model performances from the experiments in Section 4 across multiple dimensions of UNIVERSALCEFR on results for select languages that we are qualified to interpret.

**English**. Analysis of model performance shows that using fine-tuned models and linguistic feature-based classification (62% - 75%) obtains the best performance compared to prompting with instruction-tuned LLMs (19%-28%). However, these models tend to provide distinct patterns of specific CEFR labels. For the prompting setup, Gemma1, Gemma3, and EuroLLM models tend to give labels within the A1 and B1 range, while fine-tuned and feature-based models tend to lean towards the B1 and B2 range. For the pretrained and instruction-tuned models, this finding may be tied to A1 and B2 being the most common CEFR level band of most general-purpose texts found online, where the sources of the data from which these models are trained. For feature-based models, we note the potential effect of training and test data having higher instance counts for these level bands than A1, C1, and C2. Regarding model scale, upgraded versions from similar model families perform better than their previous versions, echoing previous findings in literature (Imperial and Tayyar Madabushi, 2024). This is particularly evident in Gemma3 being 12B in size and trained with massively multilingual data in

| Lang | A1 | A2 | B1 | B2 | C1 | C2 |
|---|---|---|---|---|---|---|
| EN | 173,005 | 119,335 | 59,634 | 20,746 | 7,122 | 675 |
| ES | 4577 | 4989 | 4,051 | 3,007 | 1,707 | 0 |
| DE | 273 | 13,208 | 12,996 | 346 | 108 | 308 |
| NL | 18 | 93 | 323 | 277 | 84 | 33 |
| CS | 1 | 92 | 77 | 38 | 2 | 0 |
| IT | 17 | 261 | 267 | 1 | 0 | 0 |
| FR | 106 | 302 | 404 | 335 | 210 | 98 |
| ET | 0 | 266 | 406 | 293 | 215 | 0 |
| PT | 204 | 62 | 270 | 59 | 80 | 0 |
| AR | 62 | 207 | 407 | 445 | 285 | 153 |
| HI | 203 | 219 | 223 | 203 | 182 | 145 |
| RU | 327 | 234 | 331 | 256 | 192 | 69 |
| CY | 463 | 332 | 0 | 0 | 0 | 0 |
| **Total** | **179,256** | **139,600** | **79,389** | **26,006** | **10,187** | **1,481** |

Table 8: Data statistics of UniversalCEFR-train in terms of recognized CEFR levels (A1, A2, B1, B2, C1, C2) across the 13 target languages.

| Lang | Sent | Para | Doc | Diag |
|---|---|---|---|---|
| EN | 12,826 | 409,362 | 1,837 | 0 |
| ES | 0 | 713 | 31,355 | 0 |
| DE | 26,244 | 1,033 | 5,673 | 0 |
| NL | 0 | 0 | 3,596 | 0 |
| CS | 0 | 441 | 0 | 0 |
| IT | 0 | 813 | 0 | 0 |
| FR | 1,669 | 0 | 344 | 0 |
| ET | 0 | 420 | 1,277 | 0 |
| PT | 0 | 1,423 | 0 | 0 |
| AR | 1,945 | 215 | 0 | 0 |
| HI | 1,491 | 0 | 0 | 0 |
| RU | 1,758 | 0 | 0 | 0 |
| CY | 1,107 | 109 | 41 | 115 |
| **Total** | **47,040** | **414,529** | **115** | **44,123** |

Table 9: Data statistics of UniversalCEFR-train in terms of levels (sentence, paragraph, document, dialogue) across the 13 target languages.

140+ languages and obtaining 28% in weighted F1 compared to Gemma1, which is 7B in size and English-centric, obtaining 21.8%. We note a potential *default effect* in using these models where additional specific CEFR descriptor information is not needed if the texts being evaluated are in English, due to the majority of data in the context of CEFR that is reflected in the training data being English.

**Spanish**. Fine-tuned models outperform other setups, with feature-based approaches, especially Random Forest, achieving reasonable comparative performance. Moreover, multilingual models provide noticeable performance gains when compared to the English-only model. As per prompting strategy, for smaller multilingual models the language specific prompt seems to play a role in improving the performance as it also does for the Gemma1 English-only model, however the Gemma3 with 12B parameter is not affected by this and it's been able to produce the best results of the LLMs (plus more sophisticated prompting strategies). As for the granularity of the input, models perform noticeably better at the document level than at the paragraph level, indicating that longer contexts are easier to classify than short ones. Finally, it is worth reporting a noticeable error of Gemma1: the prediction of C2 grade level, which does not exist in the Spanish dataset.

**Hindi**. Both the Gemma models perform poorly compared to the fine-tuned XLM-R and the Random Forest variants and tend to classify most Hindi test items as A1 or A2. For example, Gemma1 puts 57% of Hindi test samples as A1, whereas there are only 19% of the test samples labeled as A1 in the gold standard labels. This is in line with the general trend noticed in Section 5.2, as the Hindi subset is entirely sentence-level. The distribution is closer to the Gold distribution for the fine-tuned and feature-engineered models. XLM-R fine-tuned models give the best performance amongst all models for Hindi, both in terms of exact category prediction and in terms of the degree of error (i.e., being within 1 level above or below the correct level). Finally, we looked at the correlation between a simple approximation of text length (calculated as the number of space-separated tokens), a commonly used variable in such automated language assessment approaches in NLP research, and the CEFR gold labels, as well as model-predicted labels, after converting them to a numeric scale. There was a high correlation between text length and the gold labels (0.7), which was also seen with the XLM-R model (0.74) and the Random Forest models (0.77). However, the Gemma models only had correlations of 0.44 and 0.54, respectively, with text length. However, considering that the Hindi subset only has sentence-level annotations without a larger context, it may be challenging to achieve further consistency with the gold standard labels, given the size of the annotated dataset.

| Lang | A1 | A2 | B1 | B2 | C1 | C2 |
|---|---|---|---|---|---|---|
| EN | 19,449 | 13,151 | 6,643 | 2,384 | 797 | 85 |
| ES | 1535 | 1226 | 904 | 471 | 285 | 0 |
| DE | 32 | 2,494 | 2,392 | 60 | 13 | 41 |
| NL | 6 | 70 | 235 | 230 | 99 | 32 |
| CS | 0 | 14 | 9 | 6 | 0 | 0 |
| IT | 3 | 33 | 23 | 1 | 0 | 0 |
| FR | 13 | 30 | 39 | 43 | 20 | 12 |
| ET | 0 | 19 | 52 | 21 | 25 | 0 |
| PT | 61 | 213 | 50 | 144 | 19 | 61 |
| AR | 7 | 26 | 56 | 53 | 35 | 15 |
| HI | 22 | 30 | 20 | 16 | 12 | 13 |
| RU | 34 | 23 | 25 | 34 | 21 | 9 |
| CY | 67 | 44 | 0 | 0 | 0 | 0 |
| **Total** | **21,229** | **17,373** | **10,448** | **3,463** | **1,326** | **268** |

Table 10: Data statistics of **UniversalCEFR-dev** in terms of recognized CEFR levels (A1, A2, B1, B2, C1, C2) across the 13 target languages.

| Lang | Sent | Para | Doc | Diag |
|---|---|---|---|---|
| EN | 1,274 | 40,980 | 0 | 255 |
| ES | 0 | 51 | 0 | 4,370 |
| DE | 4,168 | 79 | 0 | 785 |
| NL | 0 | 0 | 0 | 672 |
| CS | 0 | 29 | 0 | 0 |
| IT | 0 | 60 | 0 | 0 |
| FR | 146 | 0 | 0 | 11 |
| ET | 0 | 19 | 0 | 98 |
| PT | 0 | 548 | 0 | 0 |
| AR | 188 | 4 | 0 | 0 |
| HI | 113 | 0 | 0 | 0 |
| RU | 146 | 0 | 0 | 0 |
| CY | 111 | 0 | 0 | 0 |
| **Total** | **6,146** | **41,770** | **0** | **6,191** |

Table 11: Data statistics of **UniversalCEFR-dev** in terms of levels (sentence, paragraph, document, dialogue) across the 13 target languages.

Future research should expand the available CEFR graded resources both in terms of quantity as well as granularity for the language.

**Russian**. The Russian results follow the broad patterns reported in the paper, but their rich inflectional morphology and their comparatively limited training data amplify several effects. Gemma1 (34.8%) greatly over-predicts texts as beginner-level (only 5% of texts had predictions above B1), confirming the overall trend that small, English-centric LLMs struggle most with morphologically rich languages. Gemma3 (37.4%) partially corrects this, but still massively under-predicts B2 and C2. XLM-R (49.6%) mirrors the gold distribution most faithfully, possibly because its multilingual vocabulary gives it better coverage of Russian inflectional morphology, a pattern also seen for other highly inflected languages such as Czech. The two Random Forest models (47.2% and 47.8%) under-predict A2 and C2 but otherwise match the gold shape, showing that handcrafted lexical and morpho-syntactic features capture useful Russian-specific signals even with limited data. Subword-level multilingual models (XLM-R) or explicit morpho-syntactic features (RF) are best suited to capture the meanings and relations between Russian words. Text length appears to be a false friend; although it does correlate highly with readability (r=0.65), it also appears to be the source of many errors; top-performing model outputs had text length correlations as high as 0.73. Since this experiment with Russian is limited to sentence-level readability, comparison with previous research on Russian readability assessment is not straightforward. However, the weighted F1 (49.6%) of the best-performing model (XLM-R) is below state-of-the-art results for longer texts, including 67% (Reynolds, 2016), 74% (Solnyshkina et al., 2018), and 78% (Blinova and Tarasov, 2022). Most likely, this difference is partly due to the absence of Russian-specific morphosyntactic features that have been highly informative in previous studies' models.

**Portuguese**. Comparing the different setups, we can see that the results for Portuguese follow the global tendency, with fine-tuned models achieving the highest performance, followed by feature-based models, and with prompting taking the last place. Although this study only covers paragraph-level learner data for Portuguese, similar patterns were observed on reference data (Ribeiro et al., 2024b). However, comparing the results with those of other languages and, particularly, those with paragraph-level learner data, we can see that Portuguese is the language with the lowest performance (≈33.5%). Several factors may contribute to this outcome. For instance, Portuguese is one of the languages with the least available training data, and the distribution of proficiency labels is right-skewed (especially in COPLE2). Furthermore, the data consists of texts written by learners from a wide range of

| LANG | A1 | A2 | B1 | B2 | C1 | C2 |
|---|---|---|---|---|---|---|
| EN | 107 | 114 | 132 | 129 | 83 | 35 |
| ES | 49 | 58 | 140 | 108 | 45 | 0 |
| DE | 14 | 264 | 238 | 67 | 9 | 8 |
| NL | 4 | 21 | 69 | 77 | 22 | 7 |
| CS | 0 | 82 | 79 | 37 | 2 | 0 |
| IT | 9 | 87 | 104 | 0 | 0 | 0 |
| FR | 32 | 57 | 132 | 100 | 63 | 16 |
| ET | 0 | 110 | 130 | 93 | 67 | 0 |
| PT | 49 | 50 | 47 | 30 | 13 | 11 |
| AR | 12 | 26 | 162 | 145 | 40 | 15 |
| HI | 38 | 34 | 42 | 42 | 28 | 16 |
| RU | 41 | 36 | 52 | 35 | 24 | 12 |
| CY | 233 | 232 | 0 | 0 | 0 | 0 |
| **Total** | **588** | **1,171** | **1,327** | **863** | **396** | **120** |

Table 12: Data statistics of **UNIVERSALCEFR-TEST** in terms of recognized CEFR levels (A1, A2, B1, B2, C1, C2) across the 13 target languages.

| 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 less 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

| 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 4. 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.

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.

**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 Information

## 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[10] for imageability and concreteness (Ljubešić, 2018). Most of the features have already been implemented in the `text-simplification-evaluation` (`TSEval`) package[11] (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 3.

## 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

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

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

| Field | Description |
|---|---|
| title | The unique title of the text retrieved from its original corpus (NA if there are no titles such as CEFR-assessed sentences or paragraphs). |
| lang | The source language of the text in ISO 638-1 format (e.g., en for English). |
| source_name | The source dataset name where the text is collected as indicated from their source dataset, paper, and/or documentation (e.g., cambridge-exams from Xia et al. (2016)). |
| format | The format of the text in terms of level of granularity as indicated from their source dataset, paper, and/or documentation. The recognized formats are the following: [document-level, paragraph-level, discourse-level, sentence-level]. |
| category | The classification of the text in terms of who created the material. The recognized categories are reference for texts created by experts, teachers, and language learning professionals and learner for texts written by language learners and students. |
| cefr_level | The CEFR level associated with the text. The six recognized CEFR levels are the following: [A1, A2, B1, B2, C1, C2]. A small fraction (<1%) of text in UNIVERSALCEFR contains unlabelled text, texts with plus signs (e.g., A1+), and texts with no level indicator (e.g., A, B). |
| license | The licensing information associated with the text (Unknown if not stated). |
| text | The actual content of the text itself. |

Table 15: The structured JSON fields with descriptions and examples used as the standardized uniform format for building the UNIVERSALCEFR dataset. All instances validated from the collection of CEFR-labelled corpora conform to this format.

| Category | Feature Name |
|---|---|
| Length | doc_num_sents |
| | doc_num_tokens |
| | num_characters |
| | num_characters_per_sentence |
| | num_characters_per_word |
| | num_syllables_in_sentence |
| | num_syllables_per_sentence |
| | num_syllables_per_word |
| | num_words |
| Lexical | average_pos_in_freq_table |
| | lexical_complexity_score |
| Morphosyntactic | ratio_Tense_Past |
| | ratio_of_determiners |
| | ratio_of_numerals |
| | ratio_of_pronouns |
| Psycholinguistic | concreteness |
| | imagebility |
| Readability | sentence_fkgl |
| | sentence_fre |
| Syntactic | avg_distance_between_words |
| | average_length_VP |
| | parse_tree_height |
| | ratio_of_coordinating_clauses |

Table 16: List of linguistic features occurring in the top 10 of at least three languages. We use this list for the TopFeatures subset used in the experiment result in Table 3.

| Hyperparameter | Value |
|---|---|
| Learning rate | $3.6 \times 10^{-5}$ |
| Train batch size | 2 |
| Evaluation batch size | 3 |
| Random seed | 42 |
| Gradient accumulation steps | 16 |
| Total effective batch size | 32 |
| Optimizer | `adamw_torch_fused` |
| Betas | $(0.9, 0.999)$ |
| Epsilon | $10^{-8}$ |
| Learning-rate scheduler | linear |
| Warm-up ratio | 0.1 |

Table 17: Hyperparameter values used for fine-tuning pretrained language models.

| Hyperparameter | Value |
|---|---|
| Sampling | False |
| Max New Tokens | 10 |
| Data Type | torch.bfloat16 |
| GPU | 4 x NVIDIA RTX A5000 (24GB) |

Table 18: Hyperparameter values and GPU information used for prompting instruction-tuned models.

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 A1 level, particularly for psycholinguistic features such as imageability ($\rho = 0.48$) and concreteness ($\rho = 0.46$), as well as punctuation-related measures. This suggests that certain surface-level and lexical-semantic features may be especially informative at the lowest proficiency level. A notable case is the feature of word length in characters, which shows a negative correlation at A1 ($\rho = -0.45$), becomes neutral at A2, and shifts to a positive correlation at B1 and higher levels. This pattern may reflect increasing lexical complexity with proficiency. Similarly, features related to syntactic structure, such as the ratio of past tense verbs and phrase length, generally shift from weak negative to weak positive correlations as proficiency increases, indicating progressive syntactic development. Overall, the directionality of several features suggests dynamic usage patterns across CEFR bands, even if the correlation strengths remain modest.

## F Hyperparameter Values

We detail the hyperparameter values used for fine-tuning pretrained (ModernBERT, EuroBERT, and XLM-R) and instruction-tuned language models (Gemma1, Gemma3, and EuroLLM) in Tables 17 and 18, respectively.

| Lang | Sent | Para | Doc | Overall |
|---|---|---|---|---|
| AR | **55.7** | 32.6 | - | 43.1 |
| CY | **86.9** | 72.5 | 61.5 | 72.7 |
| CS | - | 68.8 | - | 68.8 |
| DE | 65.4 | 71.1 | **83.4** | 73.2 |
| EN | 68.3 | **100.0** | 57.6 | 75.5 |
| ES | - | 40.6 | **98.0** | 69.69 |
| ET | - | **93.6** | 84.0 | 88.9 |
| FR | **57.6** | - | 44.2 | 51.7 |
| HI | 52.9 | - | - | 52.9 |
| IT | - | 83.3 | - | 83.3 |
| NL |  | - | 59.0 | 59.0 |
| PT | - | 29.2 | - | 29.2 |
| RU | 49.6 | - | - | 49.6 |

Table 19: Weighted F1 scores for the fine-tuned XLM-R (top model across all setups) performance on the UniversalCEFR-test, classified by the granularity levels of the data.

## G Additional Context on Restrictions of GDPR-Protected Datasets

The critical aspect of the GDPR is that it gives data subjects (e.g., L2 learners of CEFR) the right to withdraw their personal information from processing, which requires data processors to store both the signed consents and the ID mappings (i.e., mappings between the names of the real people and their IDs in a released corpora). As long as these documents exist and reidentification is theoretically possible, the data falls under the scope of the GDPR. Further complicating factors are national legislations and ethical regulations, such as archival laws, that treat any data produced at universities—including those used for language proficiency assessment such as essays, recorded dialogues, and written texts from personal experiences—as the property of the state (and hence making destruction of the ID mappings a non-trivial act) (European Parliament and Council, 2016).

Yet another upcoming challenge is the EU AI Act (European Parliament and Council, 2024) that implies that AI models trained on personal data should inherit the same license as the data they have been trained on, meaning that the models will be under the scope of the GDPR. We hypothesize that the non-restricted datasets included in UniversalCEFR either do not contain personal information or were collected before the GDPR, since they are already openly accessible to the public. We further hypothesize that the datasets currently under the GDPR will eventually have their ID mappings destroyed and will no longer be subject to the GDPR. This may mean that the learner corpora that can be added to UniversalCEFR will grow with time.

## H Full Dataset Directory of UniversalCEFR

We provide the complete information of qualified corpora included in the current UniversalCEFR collection to form a directory of datasets. Aside from eight per-instance information included in the standardized JSON format in Table 15, we also report five per-corpus information as listed below:

- Annotation method used (manual, computer-assisted, or NA).
- Total number of expert annotators.
- Distinct L1 learners per language for learner corpora.
- Inter-annotator agreement (IAA) metric and score.
- Reference to published paper or repository.

## I Prompt Templates

We provide the complete copies of the prompt templates used in prompting experiments with instruction-tuned LLMs as described in Section 4. The prompt templates are categorized by color based on the setup: **Base**, **En-Read**, **Lang-Read**, **En-Write**, **Lang-Write**.

## J Welsh Data Collection

One of the contributions of UniversalCEFR is the release of the first-ever open dataset for the Welsh language (CY) with gold-standard CEFR labels for A1 and A2. To obtain this data, we corresponded with data maintainers from Learn Welsh (`https://learnwelsh.cymru/`), which is a compilation of expert-created books (reference texts) and acquired PDF versions. This resource can be shared in any format for non-commercial research, which fits the goal of UniversalCEFR. We then manually extracted qualified texts according to the four levels of granularity: sentence, paragraph, dialogue, and document. The distribution of CEFR levels and text granularity for this new Welsh dataset can be found in Table 6 and 7, respectively.

| Category | Feature | Short Description | Resource | Corr. $\rho$ (avg.) | Corr. $\rho$ (SD) |
|---|---|---|---|---|---|
| discourse | ratio_referential | Ratio of referential tokens to all tokens based on dependency tree relations | SpaCy, Stanza, TSEval | 0.385 | |
| Length | doc_num_sents | Number of sentences per document / text | SpaCy, Stanza, | 0.3934 | 0.176 |
| | doc_num_tokens | Number of tokens per document / text | SpaCy, Stanza, | -0.4254 | 0.2398 |
| | num_characters | Number of characters per document / text | SpaCy, Stanza, TSEval | 0.1819 | 0.181 |
| | num_characters_per_sentence | Number of characters per sentence | SpaCy, Stanza, TSEval | 0.541 | 0.2242 |
| | num_characters_per_word | Number of characters per word | SpaCy, Stanza, TSEval | -0.3962 | 0.2414 |
| | num_sentences | Number of sentences per document / text | SpaCy, Stanza, TSEval | 0.385 | |
| | num_syllables_in_sentence | Number of syllables in document / text | SpaCy, Stanza, pyphen, TSEval | 0.219 | 0.1423 |
| | num_syllables_per_sentence | Number of syllables per sentence | SpaCy, Stanza, pyphen, TSEval | -0.685 | 0.737 |
| | num_syllables_per_word | Number of syllables per word | SpaCy, Stanza, pyphen, TSEval | 0.2394 | |
| | num_words | Number of tokens per document / text | SpaCy, Stanza, TSEval | 0.658 | 0.298 |
| | num_words_per_sentence | Number of tokens per sentence | SpaCy, Stanza, TSEval | | |
| Lexical | average_pos_in_freq_table | Average frequency rank of tokens in FastText embeddings | SpaCy, Stanza, TSEval | | |
| | lexical_complexity_score | Lexical complexity based on ranks of FastText embeddings | SpaCy, Stanza, FastText, TSEval | | |
| | type_token_ratio | Type-token-Ratio | SpaCy, Stanza, TSEval | | |
| | max_pos_in_freq_table | Maximum frequency rank of tokens in FastText embeddings | SpaCy, Stanza, TSEval | 0.1293 | |
| Psycholinguistic | concreteness | Concreteness of words based on MEGAHR and FastText-Embeddings | MEGA.HR crossling | 0.1429 | 0.139 |
| | imagebility | imagebility of words based on MEGAHR and FastText-Embeddings | MEGA.HR crossling | 0.253 | 0.1635 |
| Readability | sentence_fkgl | Flesch-Kincaid-Grading-Level, designed for English | SpaCy, Stanza, TSEval | 0.2459 | 0.1615 |
| | sentence_fre | Flesch-Reading Ease, designed for English | SpaCy, Stanza, TSEval | 0.238 | 0.1572 |
| Syntactic | avg_distance_betweeen_verb_particle | Average distance between verb and particle based on dependency tree | SpaCy, Stanza, | -0.21 | 0.1543 |
| | avg_distance_betweeen_words | Average distance between words based on dependency tree | SpaCy, Stanza, | -0.59 | 0.179 |
| | max_distance_betweeen_verb_particles | Maximum distance between verb and particle based on dependency tree | SpaCy, Stanza, | -0.5 | 0.2121 |
| | max_distance_betweeen_words | Maximum distance between words based on dependency tree | SpaCy, Stanza, | 0.143 | 0.1557 |
| | check_if_head_is_noun | Whether the head of the dependency tree is a noun | SpaCy, Stanza, TSEval | -0.89 | 0.1385 |
| | check_if_head_is_verb | Whether the head of the dependency tree is a verb | SpaCy, Stanza, TSEval | -0.2226 | 0.1621 |
| | check_if_one_child_of_root_is_subject | Whether a child of a root is a subject (not a verb) | SpaCy, Stanza, TSEval | 0.2361 | 0.1428 |
| | check_passive_voice | Whether a sentence is in passive voice | SpaCy, Stanza, TSEval | 0.542 | 0.178 |
| | average_length_NP | Average length of noun phrase in tokens | SpaCy, Stanza, TSEval | -0.954 | 0.1394 |
| | average_length_VP | Average length of verb phrase in tokens | SpaCy, Stanza, TSEval | 0.1797 | 0.1353 |
| | avg_length_PP | Average length of prepositional phrase in tokens | SpaCy, Stanza, TSEval | 0.1278 | 0.854 |
| | parse_tree_height | Depth or height of the dependency tree | SpaCy, Stanza, TSEval | 0.3262 | 0.151 |
| | ratio_clauses | Ratio of tokens associated to a clause to all tokens based on dependency tree relations | SpaCy, Stanza, TSEval | -0.3364 | 0.1627 |
| | ratio_of_coordinating_clauses | Ratio of tokens associated to a coordinating clause to all tokens based on dependency tree relations | SpaCy, Stanza, TSEval | 0.883 | 0.1329 |
| | ratio_of_subordinate_clauses | Ratio of tokens associated to a subordinating clause to all tokens based on dependency tree relations | SpaCy, Stanza, TSEval | 0.214 | 0.746 |
| | ratio_prepositional_phrases | Ratio of tokens associated to a prepositional phrase to all tokens based on dependency tree relations | SpaCy, Stanza, TSEval | 0.4738 | 0.1873 |
| | ratio_relative_phrases | Ratio of tokens associated to a relative clause to all tokens based on dependency tree relations | SpaCy, Stanza, TSEval | -0.3976 | 0.1884 |
| | is_non_projective | Whether a dependency tree is non projective | SpaCy, Stanza, TSEval | 0.51 | 0.253 |
| Morphosyntactic | ratio_Abbr_Yes | Ratio of nouns which are an abbreviation to all nouns | SpaCy, Stanza, UniversalDependencies | 0.1822 | 0.869 |
| | ratio_Case_Abe | Ratio of nouns in abessive case to all nouns | SpaCy, Stanza, UniversalDependencies | | |
| | ratio_Case_Acc | Ratio of nouns in accusative case to all nouns | SpaCy, Stanza, UniversalDependencies | | |
| | ratio_Case_Ben | Ratio of nouns in benefactive case to all nouns | SpaCy, Stanza, UniversalDependencies | | |
| | ratio_Case_Cau | Ratio of nouns in causative case to all nouns | SpaCy, Stanza, UniversalDependencies | 0.1196 | |
| | ratio_Case_Cmp | Ratio of nouns in comparative case to all nouns | SpaCy, Stanza, UniversalDependencies | 0.3528 | 0.1248 |
| | ratio_Case_Cns | Ratio of nouns in considerative case to all nouns | SpaCy, Stanza, UniversalDependencies | 0.224 | 0.224 |
| | ratio_Case_Com | Ratio of nouns in comitative case to all nouns | SpaCy, Stanza, UniversalDependencies | -0.286 | 0.133 |
| | ratio_Case_Dat | Ratio of nouns in dative case to all nouns | SpaCy, Stanza, UniversalDependencies | 0.2252 | |
| | ratio_Case_Dis | Ratio of nouns in distributive case to all nouns | SpaCy, Stanza, UniversalDependencies | | |
| | ratio_Case_Equ | Ratio of nouns in equative case to all nouns | SpaCy, Stanza, UniversalDependencies | 0.5228 | |
| | ratio_Case_Erg | Ratio of nouns in ergative case to all nouns | SpaCy, Stanza, UniversalDependencies | | |

Table 20: Overview of all 100 features, including correlation coefficient with CEFR level across all languages.

| Category | Feature | Short Description | Resource | Corr. $\rho$ (avg.) | Corr. $\rho$ (SD) |
|---|---|---|---|---|---|
| Morphosyntactic | ratio_Case_Par | Ratio of nouns in partitive case to all nouns | SpaCy, Stanza, UniversalDependencies | | |
| | ratio_Case_Ins | Ratio of nouns in instrumental case to all nouns (relevant for CZ) | SpaCy, Stanza, UniversalDependencies | 0.569 | |
| | ratio_Case_Ess | Ratio of nouns in essive case to all nouns (relevant for ET) | SpaCy, Stanza, UniversalDependencies | | |
| | ratio_Case_Gen | Ratio of nouns in genitive case to all nouns | SpaCy, Stanza, UniversalDependencies | | |
| | ratio_Case_Nom | Ratio of nouns in nominative case to all nouns | SpaCy, Stanza, UniversalDependencies | -0.59 | |
| | ratio_Case_Tem | Ratio of nouns in temporal case to all nouns | SpaCy, Stanza, UniversalDependencies | 0.2172 | 0.119 |
| | ratio_Case_Tra | Ratio of nouns in translative case to all nouns (relevant for ET) | SpaCy, Stanza, UniversalDependencies | -0.243 | 0.145 |
| | ratio_Case_Voc | Ratio of nouns in vocative case to all nouns | SpaCy, Stanza, UniversalDependencies | -0.782 | 0.12 |
| | ratio_Definite_Com | Ratio of complex nouns to all nouns (relevant for AR) | SpaCy, Stanza, UniversalDependencies | 0.798 | |
| | ratio_Definite_Cons | Ratio of nouns in construct state to all nouns (relevant for AR) | SpaCy, Stanza, UniversalDependencies | 0.115 | |
| | ratio_Definite_Def | Ratio of definite nouns to all nouns | SpaCy, Stanza, UniversalDependencies | 0.276 | 0.1269 |
| | ratio_Definite_Ind | Ratio of indefinite nouns to all nouns | SpaCy, Stanza, UniversalDependencies | 0.811 | |
| | ratio_Foreign_Yes | Ratio of nouns which are foreign to all nouns | SpaCy, Stanza, UniversalDependencies | 0.2426 | 0.1892 |
| | ratio_Mood_Cnd | Ratio of verbs with conditional mood to all verbs | SpaCy, Stanza, UniversalDependencies | 0.16 | 0.1856 |
| | ratio_Mood_Imp | Ratio of verbs with imperative mood to all verbs | SpaCy, Stanza, UniversalDependencies | 0.2287 | 0.189 |
| | ratio_Mood_Ind | Ratio of verbs with indicative mood to all verbs | SpaCy, Stanza, UniversalDependencies | 0.2199 | |
| | ratio_Mood_Jus | Ratio of verbs with jussive mood to all verbs | SpaCy, Stanza, UniversalDependencies | 0.223 | 0.727 |
| | ratio_Mood_Qot | Ratio of verbs with quotative mood to all verbs | SpaCy, Stanza, UniversalDependencies | 0.296 | 0.1296 |
| | ratio_Mood_Sub | Ratio of verbs with subjunctive mood to all verbs | SpaCy, Stanza, UniversalDependencies | 0.2782 | 0.1694 |
| | ratio_Number_Dual | Ratio of nouns in dual number to all nouns | SpaCy, Stanza, UniversalDependencies | | |
| | ratio_Number_Plur | Ratio of nouns in plural number to all nouns | SpaCy, Stanza, UniversalDependencies | -0.1113 | 0.1636 |
| | ratio_Number_Sing | Ratio of nouns in singular number to all nouns | SpaCy, Stanza, UniversalDependencies | 0.225 | 0.1849 |
| | ratio_Polarity_Neg | Ratio of negative verbs to all verbs | SpaCy, Stanza, UniversalDependencies | 0.682 | |
| | ratio_Polarity_Pos | Ratio of positive verbs to all verbs | SpaCy, Stanza, UniversalDependencies | 0.3196 | 0.178 |
| | ratio_Tense_Fut | Ratio of verbs in future tense to all verbs | SpaCy, Stanza, UniversalDependencies | -0.97 | 0.2168 |
| | ratio_Tense_Imp | Ratio of verbs in imperfect to all verbs | SpaCy, Stanza, UniversalDependencies | -0.1147 | 0.131 |
| | ratio_Tense_Past | Ratio of verbs in past tense to all verbs | SpaCy, Stanza, UniversalDependencies | 0.954 | 0.1334 |
| | ratio_Tense_Pqp | Ratio of verbs in pluferfect to all verbs | SpaCy, Stanza, UniversalDependencies | 0.144 | 0.1525 |
| | ratio_Tense_Pres | Ratio of verbs in present tenst to all verbs | SpaCy, Stanza, UniversalDependencies | | |
| | ratio_Voice_Act | Ratio of verbs in active voice to all verbs | SpaCy, Stanza, UniversalDependencies | 0.5224 | 0.1874 |
| | ratio_Voice_Mid | Ratio of verbs in middle voice to all verbs | SpaCy, Stanza, UniversalDependencies | 0.531 | 0.1656 |
| | ratio_Voice_Pass | Ratio of verbs in passive voice to all verbs | SpaCy, Stanza, UniversalDependencies | 0.3895 | 0.1613 |
| | ratio_mwes | Ratio of multi-word expressions to all tokens based on dependency tree relations | SpaCy, Stanza, TSEval | -0.324 | 0.112 |
| | ratio_named_entities | Ratio of multi-word expressions to all tokens based on SpaCy pipeline | SpaCy, Stanza, TSEval | 0.525 | 0.191 |
| | ratio_of_adjectives | Ratio of adjectives to all tokens | SpaCy, Stanza, TSEval | 0.4634 | 0.2272 |
| | ratio_of_adpositions | Ratio of adpositions to all tokens | SpaCy, Stanza, TSEval | 0.3924 | 0.1587 |
| | ratio_of_adverbs | Ratio of adverbs to all tokens | SpaCy, Stanza, TSEval | 0.479 | 0.181 |
| | ratio_of_auxiliary_verbs | Ratio of auxiliary verbs to all tokens | SpaCy, Stanza, TSEval | 0.4863 | 0.1585 |
| | ratio_of_conjunctions | Ratio of conjunctions to all tokens | SpaCy, Stanza, TSEval | 0.3485 | 0.1325 |
| | ratio_of_determiners | Ratio of determiners to all tokens | SpaCy, Stanza, TSEval | 0.4324 | 0.1558 |
| | ratio_of_function_words | Ratio of function words to all tokens based on dependency tree relations | SpaCy, Stanza, TSEval | 0.146 | 0.747 |
| | ratio_of_interjections | Ratio of interjections to all tokens | SpaCy, Stanza, TSEval | 0.649 | 0.138 |
| | ratio_of_nouns | Ratio of nouns to all tokens | SpaCy, Stanza, TSEval | 0.4657 | 0.1517 |
| | ratio_of_numerals | Ratio of numerals to all tokens | SpaCy, Stanza, TSEval | 0.1918 | 0.119 |
| | ratio_of_particles | Ratio of particles to all tokens | SpaCy, Stanza, TSEval | 0.1171 | 0.134 |
| | ratio_of_pronouns | Ratio of pronouns to all tokens | SpaCy, Stanza, TSEval | | |
| | ratio_of_punctuation | Ratio of punctuation marks to all tokens | SpaCy, Stanza, TSEval | 0.172 | 0.1362 |
| | ratio_of_symbols | Ratio of symbols to all tokens | SpaCy, Stanza, TSEval | 0.2342 | 0.1286 |
| | ratio_of_verbs | Ratio of verbs to all tokens | SpaCy, Stanza, TSEval | 0.7 | 0.1913 |
| | verb_noun_ratio | How many verbs occur per noun? The higher the value (the more verbs), the easier the text | SpaCy, Stanza, | 0.34 | 0.1253 |

Table 21: Overview of all 100 features, including correlation coefficient with CEFR level across all languages. Part II.

| Corpus Name | Lang Code (ISO 638-1) | Format | Category | Size | Annotation Method | Expert Annotators | Distinct L1 | Inter-Annotator Agreement | CEFR Coverage | License | Resource |
|---|---|---|---|---|---|---|---|---|---|---|---|
| cambridge-exams | en | document-level | reference | 331 | n/a | n/a | n/a | n/a | A1-C2 | CC BY-NC-SA 4.0 | Xia et al. (2016) |
| elg-cefr-en | en | document-level | reference | 712 | manual | 3 | n/a | n/a | A1-C2, plus | CC BY-NC-SA 4.0 | Breuker (2022) |
| cefr-sp | en | sentence-level | reference | 17,000 | manual | 2 | n/a | $r = 0.75, 0.73$ | A1-C2 | CC BY-NC-SA 4.0 | Arase et al. (2022) |
| elg-cefr-de | de | document-level | reference | 509 | manual | 3 | n/a | n/a | A1-C2 | CC BY-NC-SA 4.0 | Breuker (2022) |
| elg-cefr-nl | nl | document-level | reference | 3,596 | manual | 3 | n/a | n/a | A1-C2, plus | CC BY-NC-SA 4.0 | Breuker (2022) |
| icle500 | en | document-level | learner | 500 | manual | 28 | ur, pa, bg, zh, cs, nl, fi, fr, de, el, hu, it, ja, ko, lt, mk, no, fa, pl, pt, ru, sr, es, sv, tn, tr | Rasch $\kappa = -0.02$ | A1-C2, plus | CC BY-NC 4.0 | Thwaites et al. (2024),<br>Granger et al. (2009) |
| cefr-asag | en | paragraph-level | learner | 299 | manual | 3 | fr | Krippendorf $\alpha = 0.81$ | A1-C2 | CC BY-NC-SA 4.0 | Tack et al. (2017) |
| merlin-cs | cs | paragraph-level | learner | 441 | manual | multiple | hu, de, fr, ru, pl, en, sk, es | n/a | A2-B2 | CC BY-SA 4.0 | Boyd et al. (2014) |
| merlin-it | it | paragraph-level | learner | 813 | manual | multiple | hu, de, fr, ru, pl, en, sk, es | n/a | A1-B1 | CC BY-SA 4.0 | Boyd et al. (2014) |
| merlin-de | de | paragraph-level | learner | 1,033 | manual | multiple | hu, de, fr, ru, pl, en, sk, es | n/a | A1-C1 | CC BY-SA 4.0 | Boyd et al. (2014) |
| hablacultura | es | paragraph-level | reference | 710 | manual | multiple | n/a | n/a | A2-C1 | CC BY NC 4.0 | Vásquez-Rodríguez et al. (2022) |
| kwiziq-es | es | document-level | reference | 206 | manual | multiple | n/a | n/a | A1-C1 | CC BY NC 4.0 | Vásquez-Rodríguez et al. (2022) |
| kwiziq-fr | fr | document-level | reference | 344 | manual | multiple | n/a | n/a | A1-C1 | CC BY NC 4.0 | Original |
| caes | es | document-level | learner | 30,935 | computer-assisted | multiple | pt, zh, ar, fr, ru | n/a | A1-C1 | CC BY NC 4.0 | Vásquez-Rodríguez et al. (2022) |
| deplain-web-doc | de | document-level | reference | 394 | manual | 2 | n/a | Cohen $\kappa = 0.85$ | A1,A2,B2,C2 | CC-BY-SA-3,<br>CC-BY-4,<br>CC-BY-NC-ND-4,<br>save_use_share | Stodden et al. (2023) |
| deplain-apa-doc | de | document-level | reference | 483 | manual | 2 | n/a | Cohen $\kappa = 0.85$ | A2-B1 | CC-BY-SA-3,<br>CC-BY-4,<br>CC-BY-NC-ND-4,<br>save_use_share | Stodden et al. (2023) |
| deplain-apa-sent | de | sentence-level | reference | 483 | manual | 2 | n/a | n/a | A2-B2 | By request | Stodden et al. (2023) |
| elle | et | paragraph-level, document-level | learner | 1,697 | manual | 2 | n/a | n/a | A2-C1 | CC BY 4.0 | ALLKIVI et al. (2024),<br>Vajjala and Rama (2018) |
| efcamdat-cleaned | en | sentence-level, paragraph-level | learner | 406,062 | manual | n/a | br, zh, tw, ru, sa, mx, de, it, fr, jp, tr | n/a | A1-C1 | Cambridge | Geertzen et al. (2013),<br>Shatz (2020)<br>Huang et al. (2017) |
| beast2019-w&i | en | sentence-level | learner | 3,600 | manual | multiple | n/a | n/a | A1-C2 | Cambridge | Bryant et al. (2019),<br>Yannakoudakis et al. (2018) |
| peapl2 | pt | paragraph-level | learner | 481 | manual | n/a | zh, en, es, de, ru, fr, ja, it, nl, tet, ar, pl, ko, ro, sv | n/a | A1-C2 | CC BY SA NC 4.0 | Martins et al. (2019) |
| cople2 | pt | paragraph-level | learner | 942 | manual | n/a | zh, en, es, de, ru, fr, ja, it, nl, tet, ar, pl, ko, ro, sv | n/a | A1-C1 | CC BY SA NC 4.0 | Mendes et al. (2016) |
| zaebuc | ar | paragraph-level | learner | 214 | manual | 3 | en | Unnamed $\kappa = 0.99$ | A2-C1 | CC BY SA NC 4.0 | Habash and Palfreyman (2022) |
| readme | ar, en, fr, hi, ru | sentence-level | reference | 9,757 | computer-assisted | 2 | n/a | Krippendorf $\kappa = 0.67, 0.78$ | A1-C2 | CC BY SA NC 4.0 | Naous et al. (2024) |
| apa-lha | de | document-level | reference | 3,130 | n/a | n/a | n/a | n/a | A2-B1 | Public | Spring et al. (2021) |
| learn-welsh | cy | document-level, sentence-level, discourse-level | reference | 1,372 | manual | n/a | n/a | n/a | A1-A2 | Public | Original |

Table 22: The UNIVERSALCEFR-FULL directory of dataset information reporting full details of properties of corpora included in the main collection.

### Base CEFR prompt template

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given text or narrative and determine its CEFR level [A1, A2, B1, B2, C1, or C2] based on vocabulary complexity, grammar, and overall language proficiency. Provide only the CEFR level as output, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for reading comprehension in English (EN)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 - Learners of this level can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.

A2 - Learners of this level can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items.

B1 - Learners of this level can read straightforward factual texts on subjects related to their field of interest with a satisfactory level of comprehension.

B2 - Learners of this level can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low-frequency idioms.

C1 - Learners of this level can understand in detail lengthy, complex texts, whether or not these relate to their own area of speciality, provided they can reread difficult sections. They can also understand a wide variety of texts including literary writings, newspaper or magazine articles, and specialized academic or professional publications, provided there are opportunities for rereading and they have access to reference tools.

C2 - Learners of this level can understand virtually all types of texts including abstract, structurally complex, or highly colloquial literary and non-literary writings. They can also understand a wide range of long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

**CEFR specifications for reading comprehension in Spanish (ES)**

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Spanish** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 - Los estudiantes de este nivel pueden comprender textos muy breves y sencillos, frase por frase, recogiendo nombres, palabras y frases básicas familiares y releyendo según sea necesario.

A2 - Los estudiantes de este nivel pueden comprender textos breves y sencillos que contienen el vocabulario de mayor frecuencia, incluyendo una proporción de vocabulario internacional compartido.

B1 - Los estudiantes de este nivel pueden leer textos factuales sencillos sobre temas relacionados con su área de interés con un nivel de comprensión satisfactorio.

B2 - Los estudiantes de este nivel pueden leer con un alto grado de independencia, adaptando el estilo y la velocidad de lectura a diferentes textos y propósitos, y utilizando selectivamente las fuentes de referencia adecuadas. Poseen un amplio vocabulario de lectura activa, pero pueden tener alguna dificultad con expresiones idiomáticas de baja frecuencia.

C1 - Los estudiantes de este nivel pueden comprender con detalle textos extensos y complejos, independientemente de si se relacionan con su área de especialidad, siempre que puedan releer las secciones difíciles. También pueden comprender una amplia variedad de textos, incluyendo escritos literarios, artículos de periódicos o revistas, y publicaciones académicas o profesionales especializadas, siempre que tengan la oportunidad de releer y acceso a recursos de referencia.

C2 - Los estudiantes de este nivel pueden comprender prácticamente todo tipo de textos, incluyendo textos literarios y no literarios abstractos, estructuralmente complejos o muy coloquiales. También pueden comprender una amplia gama de textos largos y complejos, apreciando las sutiles diferencias de estilo y el significado, tanto implícito como explícito.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «`TEXT`»

Answer:

**CEFR specifications for reading comprehension in German (DE)**

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **German** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 – Lernende dieser Stufe können sehr kurze, einfache Texte Satz für Satz verstehen, indem sie bekannte Namen, Wörter und einfache Sätze aufgreifen und bei Bedarf wiederholt lesen.

A2 – Lernende dieser Stufe können kurze, einfache Texte mit dem häufigsten Wortschatz verstehen, darunter auch einen Anteil an international verbreiteten Vokabeln.

B1 – Lernende dieser Stufe können einfache Sachtexte zu Themen ihres Interessengebiets mit zufriedenstellendem Verständnis lesen.

B2 – Lernende dieser Stufe können weitgehend selbstständig lesen, indem sie Stil und Geschwindigkeit an unterschiedliche Texte und Zwecke anpassen und geeignete Referenzquellen selektiv nutzen. Sie verfügen über einen breiten aktiven Lesewortschatz, haben aber möglicherweise Schwierigkeiten mit seltenen Redewendungen.

C1 – Lernende dieser Stufe können längere, komplexe Texte detailliert verstehen, unabhängig davon, ob sie zu ihrem Fachgebiet gehören oder nicht, sofern sie schwierige Abschnitte wiederholt lesen können. Sie können außerdem eine Vielzahl von Texten verstehen, darunter literarische Schriften, Zeitungs- und Zeitschriftenartikel sowie wissenschaftliche oder professionelle Fachpublikationen, sofern Möglichkeiten zum Nachlesen bestehen und sie Zugang zu Nachschlagewerken haben.

C2 – Lernende dieser Stufe können nahezu alle Textarten verstehen, darunter abstrakte, strukturell komplexe oder stark umgangssprachliche literarische und nicht-literarische Texte. Sie können außerdem eine breite Palette langer und komplexer Texte verstehen und dabei subtile Stilunterschiede sowie implizite und explizite Bedeutungen wahrnehmen.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

**CEFR specifications for reading comprehension in Dutch (NL)**

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Dutch** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 - Leerlingen van dit niveau kunnen zeer korte, eenvoudige teksten begrijpen, één zin tegelijk, bekende namen, woorden en basiszinnen oppikken en indien nodig herlezen.

A2 - Leerlingen van dit niveau kunnen korte, eenvoudige teksten begrijpen die de meest frequente woordenschat bevatten, inclusief een deel van de gedeelde internationale woordenschatitems.

B1 - Leerlingen van dit niveau kunnen eenvoudige feitelijke teksten lezen over onderwerpen die verband houden met hun interessegebied met een bevredigend niveau van begrip.

B2 - Leerlingen van dit niveau kunnen met een grote mate van onafhankelijkheid lezen, de stijl en leessnelheid aanpassen aan verschillende teksten en doeleinden, en selectief gebruikmaken van geschikte referentiebronnen. Heeft een brede actieve leeswoordenschat, maar kan enige moeite hebben met laagfrequente idiomen.

C1 - Leerlingen van dit niveau kunnen lange, complexe teksten gedetailleerd begrijpen, ongeacht of deze betrekking hebben op hun eigen vakgebied, op voorwaarde dat ze moeilijke secties kunnen herlezen. Ze kunnen ook een breed scala aan teksten begrijpen, waaronder literaire geschriften, kranten- of tijdschriftartikelen en gespecialiseerde academische of professionele publicaties, mits er mogelijkheden zijn om ze opnieuw te lezen en ze toegang hebben tot referentietools.

C2 - Cursisten van dit niveau kunnen vrijwel alle soorten teksten begrijpen, waaronder abstracte, structureel complexe of zeer informele literaire en niet-literaire geschriften. Ze kunnen ook een breed scala aan lange en complexe teksten begrijpen, waarbij ze subtiele verschillen in stijl en impliciete en expliciete betekenis waarderen.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for reading comprehension in Czech (CS)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Czech** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 – Studenti této úrovně dokážou porozumět velmi krátkým jednoduchým textům po jedné frázi, pochytají známá jména, slova a základní fráze a přečtou si je podle potřeby.

A2 – Studenti této úrovně dokážou porozumět krátkým jednoduchým textům obsahujícím nejfrekventovanější slovní zásobu, včetně části sdílené mezinárodní slovní zásoby.

B1 – Studenti této úrovně dokážou číst přímočaré věcné texty na témata související s jejich oblastí zájmu s uspokojivou úrovní porozumění.

B2 – Studenti této úrovně dokážou číst s velkou mírou nezávislosti, přizpůsobují styl a rychlost čtení různým textům a účelům a selektivně používají vhodné referenční zdroje. Má širokou slovní zásobu aktivního čtení, ale může mít potíže s nízkofrekvenčními idiomy.

C1 – Studenti této úrovně dokážou podrobně porozumět dlouhým a složitým textům, ať už se týkají nebo netýkají jejich vlastní oblasti specializace, za předpokladu, že dokážou znovu přečíst obtížné části. Mohou také porozumět široké škále textů, včetně literárních textů, článků v novinách nebo časopisech a specializovaných akademických nebo odborných publikací, za předpokladu, že mají příležitosti k opakovanému čtení a mají přístup k referenčním nástrojům.

C2 – Studenti této úrovně mohou porozumět prakticky všem typům textů včetně abstraktních, strukturálně složitých nebo vysoce hovorových literárních a neliterárních spisů. Dokážou také porozumět široké škále dlouhých a složitých textů, ocenit jemné rozdíly ve stylu a implicitní i explicitní význam.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for reading comprehension in Italian (IT)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Italian** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 - Gli studenti di questo livello riescono a comprendere testi molto brevi e semplici, una frase alla volta, cogliendo nomi familiari, parole e frasi di base e rileggendo quando necessario.

A2 - Gli studenti di questo livello riescono a comprendere testi brevi e semplici contenenti il vocabolario più frequente, inclusa una parte di elementi di vocabolario internazionale condiviso.

B1 - Gli studenti di questo livello riescono a leggere testi fattuali semplici su argomenti correlati al loro campo di interesse con un livello di comprensione soddisfacente.

B2 - Gli studenti di questo livello riescono a leggere con un ampio grado di indipendenza, adattando stile e velocità di lettura a testi e scopi diversi e utilizzando fonti di riferimento appropriate in modo selettivo. Ha un ampio vocabolario di lettura attiva, ma può avere qualche difficoltà con idiomi a bassa frequenza.

C1 - Gli studenti di questo livello riescono a comprendere in dettaglio testi lunghi e complessi, indipendentemente dal fatto che siano correlati o meno alla propria area di specializzazione, a condizione che riescano a rileggere sezioni difficili. Possono anche comprendere un'ampia varietà di testi, tra cui scritti letterari, articoli di giornali o riviste e pubblicazioni accademiche o professionali specializzate, a condizione che vi siano opportunità di rilettura e abbiano accesso a strumenti di riferimento.

C2 - Gli studenti di questo livello possono comprendere praticamente tutti i tipi di testi, tra cui scritti letterari e non letterari astratti, strutturalmente complessi o altamente colloquiali. Possono anche comprendere un'ampia gamma di testi lunghi e complessi, apprezzando sottili distinzioni di stile e significato implicito ed esplicito.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for reading comprehension in French (FR)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **French** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 - Les apprenants de ce niveau peuvent comprendre des textes très courts et simples, phrase par phrase, en reprenant des noms, des mots et des expressions de base familiers et en les relisant si nécessaire.

A2 - Les apprenants de ce niveau peuvent comprendre des textes courts et simples contenant le vocabulaire le plus courant, y compris une partie du vocabulaire international commun.

B1 - Les apprenants de ce niveau peuvent lire des textes factuels simples sur des sujets liés à leur domaine d'intérêt avec un niveau de compréhension satisfaisant.

B2 - Les apprenants de ce niveau peuvent lire avec une grande autonomie, en adaptant leur style et leur vitesse de lecture à différents textes et objectifs, et en utilisant sélectivement des sources de référence appropriées. Possède un vocabulaire de lecture actif et étendu, mais peut éprouver des difficultés avec les expressions idiomatiques peu fréquentes.

C1 - Les apprenants de ce niveau peuvent comprendre en détail des textes longs et complexes, qu'ils relèvent ou non de leur domaine de spécialité, à condition de pouvoir relire les passages difficiles. Ils peuvent également comprendre une grande variété de textes, notamment des écrits littéraires, des articles de journaux ou de magazines, ainsi que des publications universitaires ou professionnelles spécialisées, à condition de disposer d'opportunités de relecture et d'outils de référence.

C2 - Les apprenants de ce niveau peuvent comprendre pratiquement tous les types de textes, y compris les écrits littéraires et non littéraires abstraits, structurellement complexes ou très familiers. Ils peuvent également comprendre un large éventail de textes longs et complexes, en appréciant les subtilités stylistiques et le sens implicite et explicite.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for reading comprehension in Estonian (ET)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Estonian** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 – selle taseme õppijad saavad aru väga lühikestest lihtsatest tekstidest ühe fraasi kaupa, korjavad üles tuttavad nimed, sõnad ja põhifraasid ning loevad vajaduse korral uuesti läbi.

A2 – selle taseme õppijad saavad aru lühikestest lihtsatest tekstidest, mis sisaldavad kõige sagedamini kasutatavat sõnavara, sealhulgas osa jagatud rahvusvahelistest sõnavaraüksustest.

B1 – selle taseme õppijad oskavad rahuldaval mõistustasemel lugeda otsekoheseid faktitekste nende huvivaldkonnaga seotud teemadel.

B2 – selle taseme õppijad oskavad lugeda suurel määral iseseisvalt, kohandades lugemisstiili ja -kiirust erinevate tekstide ja eesmärkidega ning kasutades valikuliselt sobivaid viiteallikaid. Tal on lai aktiivse lugemise sõnavara, kuid tal võib esineda raskusi madala sagedusega idioomidega.

C1 – selle taseme õppijad saavad üksikasjalikult aru pikkadest ja keerukatest tekstidest, olenemata sellest, kas need on seotud nende enda erialaga või mitte, eeldusel, et nad suudavad raskeid lõike uuesti lugeda. Nad saavad aru ka paljudest erinevatest tekstidest, sealhulgas kirjanduslikest kirjutistest, ajalehtede või ajakirjade artiklitest ning erialastest akadeemilistest või erialastest väljaannetest, eeldusel, et neil on võimalus uuesti lugeda ja neil on juurdepääs viitevahenditele.

C2 – selle taseme õppijad saavad aru peaaegu igat tüüpi tekstidest, sealhulgas abstraktsetest, struktuurselt keerukatest või väga kõnekeelsetest kirjanduslikest ja mittekirjanduslikest kirjutistest. Samuti saavad nad aru paljudest pikkadest ja keerulistest tekstidest, mõistes peent stiilieritlust ning kaudset ja selgesõnalist tähendust.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for reading comprehension in Portuguese (PT)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Portuguese** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 - Os alunos deste nível podem entender textos muito curtos e simples, uma única frase de cada vez, pegando nomes, palavras e frases básicas familiares e relendo conforme necessário.

A2 - Os alunos deste nível podem entender textos curtos e simples contendo o vocabulário de maior frequência, incluindo uma proporção de itens de vocabulário internacional compartilhados.

B1 - Os alunos deste nível podem ler textos factuais diretos sobre assuntos relacionados ao seu campo de interesse com um nível satisfatório de compreensão.

B2 - Os alunos deste nível podem ler com um alto grau de independência, adaptando o estilo e a velocidade de leitura a diferentes textos e propósitos, e usando fontes de referência apropriadas seletivamente. Tem um amplo vocabulário de leitura ativa, mas pode ter alguma dificuldade com expressões idiomáticas de baixa frequência.

C1 - Os alunos deste nível podem entender em detalhes textos longos e complexos, estejam eles relacionados ou não à sua própria área de especialidade, desde que possam reler seções difíceis. Eles também podem entender uma grande variedade de textos, incluindo escritos literários, artigos de jornais ou revistas e publicações acadêmicas ou profissionais especializadas, desde que haja oportunidades de releitura e tenham acesso a ferramentas de referência.

C2 - Alunos deste nível podem entender virtualmente todos os tipos de textos, incluindo escritos abstratos, estruturalmente complexos ou altamente coloquiais, literários e não literários. Eles também podem entender uma grande variedade de textos longos e complexos, apreciando sutis distinções de estilo e significado implícito e explícito.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

**CEFR specifications for reading comprehension in Arabic (ar)**

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Arabic** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 – يمكن للمتعلمين في هذا المستوى فهم نصوص قصيرة جدًا وبسيطة، جملة واحدة في كل مرة، والتقاط الأسماء والكلمات والعبارات الأساسية المألوفة، وإعادة القراءة عند الحاجة.

A2 – يمكن للمتعلمين في هذا المستوى فهم نصوص قصيرة وبسيطة تحتوي على أكثر المفردات شيوعًا، بما في ذلك جزء من المفردات الدولية المشتركة.

B1 – يمكن للمتعلمين في هذا المستوى قراءة نصوص واقعية مباشرة حول مواضيع مرتبطة بمجال اهتمامهم بمستوى مرضٍ من الفهم.

B2 – يمكن للمتعلمين في هذا المستوى القراءة بدرجة عالية من الاستقلالية، وتكييف أسلوب وسرعة القراءة وفقًا لاختلاف النصوص والأغراض، واستخدام مصادر المراجع المناسبة بشكل انتقائي. لديهم مفردات قراءة نشطة واسعة، لكن قد يواجهون بعض الصعوبة مع التعابير الاصطلاحية نادرة الاستخدام.

C1 – يمكن للمتعلمين في هذا المستوى فهم نصوص طويلة ومعقدة بالتفصيل، سواء كانت مرتبطة بمجال تخصصهم أم لا، بشرط أن يكونوا قادرين على إعادة قراءة المقاطع الصعبة. يمكنهم أيضًا فهم مجموعة واسعة من النصوص، بما في ذلك الأعمال الأدبية، ومقالات الصحف أو المجلات، والمنشورات الأكاديمية أو المهنية المتخصصة، بشرط توفر فرص لإعادة القراءة والوصول إلى أدوات المراجع.

C2 – يمكن للمتعلمين في هذا المستوى فهم جميع أنواع النصوص تقريبًا، بما في ذلك الكتابات الأدبية وغير الأدبية التي تكون مجردة أو معقدة من حيث البنية أو شديدة العامية. يمكنهم أيضًا فهم مجموعة كبيرة من النصوص الطويلة والمعقدة، وتقدير الفروقات الدقيقة في الأسلوب والمعاني الضمنية والصريحة.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for reading comprehension in Hindi (HI)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Hindi** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 - इस स्तर के शिक्षार्थी बहुत छोटे, सरल पाठों को एक बार में एक ही वाक्यांश समझ सकते हैं, परिचित नाम, शब्द और बुनियादी वाक्यांशों को चुन सकते हैं और आवश्यकतानुसार उन्हें फिर से पढ़ सकते हैं।

A2 - इस स्तर के शिक्षार्थी सबसे अधिक आवृत्ति शब्दावली वाले छोटे, सरल पाठों को समझ सकते हैं, जिसमें साझा अंतरराष्ट्रीय शब्दावली आइटम का अनुपात शामिल है।

B1 - इस स्तर के शिक्षार्थी अपनी रुचि के क्षेत्र से संबंधित विषयों पर सीधे तथ्यात्मक पाठों को संतोषजनक स्तर की समझ के साथ पढ़ सकते हैं।

B2 - इस स्तर के शिक्षार्थी बहुत हद तक स्वतंत्रता के साथ पढ़ सकते हैं, अलग–अलग पाठों और उद्देश्यों के लिए पढ़ने की शैली और गति को अनुकूलित कर सकते हैं, और उचित संदर्भ स्रोतों का चयन करके उपयोग कर सकते हैं। एक व्यापक सक्रिय पढ़ने की शब्दावली है, लेकिन कम आवृत्ति वाले मुहावरों के साथ कुछ कठिनाई का अनुभव कर सकते हैं।

C1 - इस स्तर के शिक्षार्थी लंबे, जटिल पाठों को विस्तार से समझ सकते हैं, चाहे वे उनके अपने विशेषज्ञता के क्षेत्र से संबंधित हों या नहीं, बशर्ते वे कठिन खंडों को फिर से पढ़ सकें। वे साहित्यिक लेखन, समाचार पत्र या पत्रिका लेख, और विशेष शैक्षणिक या व्यावसायिक प्रकाशनों सहित विभिन्न प्रकार के पाठों को भी समझ सकते हैं, बशर्ते उन्हें दोबारा पढ़ने के अवसर हों और उनके पास संदर्भ उपकरणों तक पहुँच हो।

C2 - इस स्तर के शिक्षार्थी लगभग सभी प्रकार के पाठों को समझ सकते हैं, जिसमें सारगर्भित, संरचनात्मक रूप से जटिल या अत्यधिक बोलचाल की साहित्यिक और गैर–साहित्यिक लेखन शामिल हैं। वे लंबे और जटिल पाठों की एक विस्तृत श्रृंखला को भी समझ सकते हैं, शैली और निहित और साथ ही स्पष्ट अर्थ के सूक्ष्म अंतरों की सराहना करते हैं।

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for reading comprehension in Russian (RU)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Russian** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 — учащиеся этого уровня могут понимать очень короткие, простые тексты по одной фразе за раз, подбирая знакомые имена, слова и основные фразы и перечитывая их по мере необходимости.

A2 — учащиеся этого уровня могут понимать короткие, простые тексты, содержащие наиболее частотную лексику, включая часть общих международных словарных единиц.

B1 — учащиеся этого уровня могут читать простые фактические тексты по темам, связанным с их областью интересов, с удовлетворительным уровнем понимания.

B2 — учащиеся этого уровня могут читать с большой степенью независимости, адаптируя стиль и скорость чтения к различным текстам и целям и выборочно используя соответствующие справочные источники. Имеет широкий активный словарный запас чтения, но может испытывать некоторые трудности с редко встречающимися идиомами.

C1 — учащиеся этого уровня могут понимать в деталях длинные, сложные тексты, независимо от того, относятся ли они к их собственной области специализации, при условии, что они могут перечитывать сложные разделы. Они также могут понимать широкий спектр текстов, включая литературные произведения, газетные или журнальные статьи, а также специализированные академические или профессиональные публикации, при условии, что есть возможности для перечитывания и у них есть доступ к справочным материалам.

C2 - Учащиеся этого уровня могут понимать практически все типы текстов, включая абстрактные, структурно сложные или очень разговорные литературные и нелитературные произведения. Они также могут понимать широкий спектр длинных и сложных текстов, оценивая тонкие различия стиля и неявного, а также явного значения.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for reading comprehension in Welsh (CY)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Welsh** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 - Gall dysgwyr y lefel hon ddeall testunau byr iawn, syml un cymal ar y tro, gan godi enwau, geiriau ac ymadroddion sylfaenol cyfarwydd ac ailddarllen yn ôl yr angen.

A2 - Gall dysgwyr y lefel hon ddeall testunau byr, syml sy'n cynnwys yr eirfa fwyaf aml, gan gynnwys cyfran o eitemau geirfa ryngwladol a rennir.

B1 - Gall dysgwyr y lefel hon ddarllen testunau ffeithiol syml ar bynciau sy'n ymwneud â'u maes diddordeb gyda lefel foddhaol o ddealltwriaeth.

B2 - Gall dysgwyr y lefel hon ddarllen yn annibynnol iawn, gan addasu arddull a chyflymder darllen i wahanol destunau a dibenion, a defnyddio ffynonellau cyfeirio priodol yn ddetholus. Yn meddu ar eirfa ddarllen weithredol eang, ond gall brofi peth anhawster gydag idiomau amledd isel.

C1 - Gall dysgwyr y lefel hon ddeall yn fanwl destunau hir a chymhleth, p'un a yw'r rhain yn ymwneud â'u maes arbenigedd eu hunain ai peidio, ar yr amod eu bod yn gallu ailddarllen adrannau anodd. Gallant hefyd ddeall amrywiaeth eang o destunau gan gynnwys ysgrifau llenyddol, erthyglau papur newydd neu gylchgronau, a chyhoeddiadau academaidd neu broffesiynol arbenigol, ar yr amod bod cyfleoedd i'w hail-ddarllen a bod offer cyfeirio ar gael iddynt.

C2 - Gall dysgwyr ar y lefel hon ddeall bron bob math o destunau gan gynnwys ysgrifau llenyddol ac anllenyddol haniaethol, strwythurol gymhleth, neu ysgrifau llenyddol ac anllenyddol hynod lafar. Gallant hefyd ddeall ystod eang o destunau hir a chymhleth, gan werthfawrogi gwahaniaethau cynnil o ran arddull ac ystyr ymhlyg yn ogystal ag ystyr amlwg.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for written production in English (EN)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 - Learners of this level can give information about matters of personal relevance (e.g. likes and dislikes, family, pets) using simple words/signs and basic expressions. Learners can also produce simple isolated phrases and sentences.

A2 - Learners of this level can produce a series of simple phrases and sentences linked with simple connectors like “and”, “but” and “because”. Learners have sufficient vocabulary for the expression of basic communicative needs and for coping with simple survival needs.

B1 - Learners of this level can produce straightforward connected texts on a range of familiar subjects within their field of interest, by linking a series of shorter discrete elements into a linear sequence. Learners have a good range of vocabulary related to familiar topics and everyday situations.

B2 - Learners of this level can produce clear, detailed texts on a variety of subjects related to their field of interest, synthesising and evaluating information and arguments from a number of sources. Learners have a good range of vocabulary for matters connected to their field and most general topics.

C1 - Learners of this level can produce clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion. Learners can alsoemploy the structure and conventions of a variety of genres, varying the tone, style and register according to addressee, text type and theme.

C2 - Learners of this level can produce clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader identify significant points. Learners have a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for written production in Spanish (ES)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Spanish** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 – Los estudiantes de este nivel pueden dar información sobre asuntos de relevancia personal (por ejemplo, gustos y disgustos, familia, mascotas) utilizando palabras/signos simples y expresiones básicas. También pueden producir frases y oraciones simples y aisladas.

A2 – Los estudiantes de este nivel pueden producir una serie de frases y oraciones simples conectadas mediante conectores sencillos como "y", "pero" y "porque". Tienen un vocabulario suficiente para expresar necesidades comunicativas básicas y afrontar necesidades simples de supervivencia.

B1 – Los estudiantes de este nivel pueden producir textos conectados de forma sencilla sobre una variedad de temas conocidos dentro de su campo de interés, enlazando una serie de elementos breves y discretos en una secuencia lineal. Tienen un buen dominio del vocabulario relacionado con temas familiares y situaciones cotidianas.

B2 – Los estudiantes de este nivel pueden producir textos claros y detallados sobre diversos temas relacionados con su campo de interés, sintetizando y evaluando información y argumentos de diversas fuentes. Poseen un buen dominio del vocabulario relacionado con su campo y con la mayoría de los temas generales.

C1 – Los estudiantes de este nivel pueden producir textos claros y bien estructurados sobre temas complejos, resaltando los aspectos relevantes, desarrollando y respaldando puntos de vista de forma extensa con puntos secundarios, razones y ejemplos pertinentes, y concluyendo adecuadamente. También pueden utilizar la estructura y convenciones de una variedad de géneros, variando el tono, el estilo y el registro según el destinatario, el tipo de texto y el tema.

C2 – Los estudiantes de este nivel pueden producir textos claros, fluidos y complejos en un estilo apropiado y efectivo, con una estructura lógica que ayuda al lector a identificar los puntos significativos. Tienen un buen dominio de un repertorio léxico muy amplio, incluyendo expresiones idiomáticas y coloquialismos; muestran conciencia de los niveles connotativos del significado.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for written production in German (DE)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **German** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 – Lernende auf diesem Niveau können Informationen zu persönlich relevanten Themen (z.B. Vorlieben und Abneigungen, Familie, Haustiere) mit einfachen Wörtern/Gebärden und grundlegenden Ausdrücken geben. Sie können auch einfache, isolierte Sätze und Wendungen produzieren.

A2 – Lernende auf diesem Niveau können eine Reihe einfacher Sätze und Wendungen bilden, die mit einfachen Konnektoren wie „und“, „aber“ und „weil“ verbunden sind. Sie verfügen über einen ausreichenden Wortschatz, um grundlegende kommunikative Bedürfnisse und einfache Überlebensbedürfnisse auszudrücken.

B1 – Lernende auf diesem Niveau können einfache, zusammenhängende Texte zu vertrauten Themen aus ihrem Interessengebiet verfassen, indem sie eine Reihe kürzerer, einzelner Elemente zu einer linearen Folge verknüpfen. Sie verfügen über einen guten Wortschatz zu bekannten Themen und Alltagssituationen.

B2 – Lernende auf diesem Niveau können klare, detaillierte Texte zu verschiedenen Themen ihres Interessengebiets verfassen, Informationen und Argumente aus mehreren Quellen zusammenfassen und bewerten. Sie verfügen über einen guten Wortschatz für Themen ihres Fachgebiets sowie für die meisten allgemeinen Themen.

C1 – Lernende auf diesem Niveau können klare, gut strukturierte Texte zu komplexen Themen verfassen, die wesentlichen Punkte herausarbeiten, Standpunkte ausführlich mit Nebenaspekten, Begründungen und passenden Beispielen untermauern und mit einem geeigneten Schluss abrunden. Sie können außerdem Aufbau und Konventionen verschiedener Textsorten anwenden und Ton, Stil und Register je nach Adressat, Texttyp und Thema variieren.

C2 – Lernende auf diesem Niveau können klare, flüssige und komplexe Texte in einem angemessenen und wirkungsvollen Stil mit einer logischen Struktur verfassen, die dem Leser hilft, wichtige Punkte zu erkennen. Sie beherrschen ein sehr breites Spektrum an Wortschatz, einschließlich idiomatischer Wendungen und umgangssprachlicher Ausdrücke, und zeigen ein Bewusstsein für konnotative Bedeutungsebenen.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

**CEFR specifications for written production in Dutch (NL)**

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Dutch** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 – Leerlingen op dit niveau kunnen informatie geven over onderwerpen van persoonlijk belang (bijv. voorkeuren en afkeuren, familie, huisdieren) met eenvoudige woorden/tekens en basisuitdrukkingen. Ze kunnen ook eenvoudige, op zichzelf staande zinnen en uitdrukkingen produceren.

A2 – Leerlingen op dit niveau kunnen een reeks eenvoudige zinnen en uitdrukkingen produceren die verbonden zijn met eenvoudige voegwoorden zoals “en”, “maar” en “omdat”. Ze beschikken over voldoende woordenschat om basisbehoeften in communicatie en eenvoudige overlevingssituaties aan te kunnen.

B1 – Leerlingen op dit niveau kunnen eenvoudige, samenhangende teksten produceren over een reeks vertrouwde onderwerpen binnen hun interessegebied, door een reeks korte, afzonderlijke elementen in een lineaire volgorde te verbinden. Ze beschikken over een goede woordenschat met betrekking tot vertrouwde onderwerpen en alledaagse situaties.

B2 – Leerlingen op dit niveau kunnen duidelijke, gedetailleerde teksten produceren over uiteenlopende onderwerpen die verband houden met hun interessegebied, waarbij ze informatie en argumenten uit meerdere bronnen synthetiseren en evalueren. Ze hebben een goede woordenschat voor onderwerpen binnen hun vakgebied en de meeste algemene thema’s.

C1 – Leerlingen op dit niveau kunnen duidelijke, goed gestructureerde teksten produceren over complexe onderwerpen, waarbij ze relevante kernpunten onderstrepen, standpunten uitgebreid onderbouwen met nevenpunten, redenen en relevante voorbeelden, en afsluiten met een passende conclusie. Ze kunnen ook de structuur en conventies van verschillende tekstgenres hanteren en toon, stijl en register aanpassen aan de ontvanger, het teksttype en het thema.

C2 – Leerlingen op dit niveau kunnen duidelijke, vloeiende en complexe teksten produceren in een gepaste en effectieve stijl, met een logische structuur die de lezer helpt belangrijke punten te identificeren. Ze beheersen een zeer uitgebreide woordenschat, inclusief idiomatische uitdrukkingen en omgangstaal, en tonen bewustzijn van connotatieve betekenislagen.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for written production in Czech (CS)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Czech** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 – Učící se na této úrovni dokážou poskytovat informace o osobně relevantních záležitostech (např. co mají a nemají rádi, rodina, domácí mazlíčci) pomocí jednoduchých slov/znaků a základních výrazů. Dokážou také vytvářet jednoduché izolované fráze a věty.

A2 – Učící se na této úrovni dokážou vytvářet sérii jednoduchých frází a vět spojených jednoduchými spojkami jako „a“, „ale“ a „protože“. Mají dostatečnou slovní zásobu pro vyjádření základních komunikačních potřeb a zvládání jednoduchých situací nutných pro přežití.

B1 – Učící se na této úrovni dokážou vytvářet přímočaré souvislé texty na řadu známých témat v rámci svého zájmového okruhu, a to spojením série kratších oddělených prvků do lineární posloupnosti. Mají dobrý rozsah slovní zásoby týkající se známých témat a každodenních situací.
B2 – Učící se na této úrovni dokážou vytvářet jasné a podrobné texty o různých tématech souvisejících s jejich oblastí zájmu, přičemž syntetizují a hodnotí informace a argumenty z různých zdrojů. Mají dobrý rozsah slovní zásoby pro témata související s jejich oborem a většinou obecných témat.

C1 – Učící se na této úrovni dokážou vytvářet jasné a dobře strukturované texty o složitých tématech, zdůrazňují důležité body, rozvíjejí a podporují názory rozsáhlým způsobem pomocí vedlejších myšlenek, důvodů a relevantních příkladů a zakončují je vhodným závěrem. Také dokážou využívat strukturu a konvence různých žánrů a měnit tón, styl a formálnost podle adresáta, typu textu a tématu.

C2 – Učící se na této úrovni dokážou vytvářet jasné, plynulé a složité texty vhodným a efektivním stylem a logickou strukturou, která pomáhá čtenáři rozpoznat důležité body. Mají výbornou znalost velmi široké slovní zásoby včetně idiomů a hovorových výrazů; projevují citlivost na konotace a jemné významové odstíny.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for written production in Italian (IT)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Italian** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 – Gli apprendenti di questo livello possono fornire informazioni su argomenti di rilevanza personale (ad esempio, gusti e preferenze, famiglia, animali domestici) utilizzando parole/segnali semplici ed espressioni di base. Possono anche produrre frasi ed enunciati semplici e isolati.

A2 – Gli apprendenti di questo livello possono produrre una serie di frasi ed enunciati semplici collegati con connettivi basilari come "e", "ma" e "perché". Possiedono un vocabolario sufficiente per esprimere bisogni comunicativi di base e per affrontare necessità semplici di sopravvivenza.

B1 – Gli apprendenti di questo livello possono produrre testi semplici e coerenti su una gamma di argomenti familiari all'interno del proprio campo di interesse, collegando una serie di elementi più brevi in una sequenza lineare. Possiedono un buon repertorio di vocaboli relativi a temi familiari e situazioni quotidiane.

B2 – Gli apprendenti di questo livello possono produrre testi chiari e dettagliati su vari argomenti legati al proprio campo di interesse, sintetizzando e valutando informazioni e argomentazioni provenienti da diverse fonti. Hanno un buon vocabolario per trattare argomenti del proprio ambito e la maggior parte dei temi generali.

C1 – Gli apprendenti di questo livello possono produrre testi chiari e ben strutturati su argomenti complessi, evidenziando le questioni salienti, sviluppando e sostenendo opinioni in modo articolato con punti secondari, motivazioni ed esempi rilevanti, e concludendo con una chiusura appropriata. Sono anche in grado di adottare la struttura e le convenzioni di diversi generi, variando tono, stile e registro in base al destinatario, al tipo di testo e al tema.

C2 – Gli apprendenti di questo livello possono produrre testi chiari, scorrevoli e complessi in uno stile appropriato ed efficace, con una struttura logica che aiuta il lettore a identificare i punti significativi. Possiedono un'ottima padronanza di un ampio repertorio lessicale, incluse espressioni idiomatiche e colloquiali, e mostrano consapevolezza dei livelli connotativi del significato.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for written production in French (FR)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **French** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 – Les apprenants de ce niveau peuvent fournir des informations sur des sujets personnels (par exemple, goûts et dégoûts, famille, animaux de compagnie) en utilisant des mots/signes simples et des expressions de base. Ils peuvent également produire des phrases et expressions simples et isolées.

A2 – Les apprenants de ce niveau peuvent produire une série de phrases et d'expressions simples reliées par des connecteurs simples comme « et », « mais » et « parce que ». Ils possèdent un vocabulaire suffisant pour exprimer des besoins communicatifs de base et faire face à des situations simples de survie.

B1 – Les apprenants de ce niveau peuvent produire des textes clairs et cohérents sur une variété de sujets familiers dans leur domaine d'intérêt, en reliant une série d'éléments plus courts dans une séquence linéaire. Ils disposent d'un bon éventail de vocabulaire lié aux sujets familiers et aux situations de la vie quotidienne.

B2 – Les apprenants de ce niveau peuvent produire des textes clairs et détaillés sur divers sujets liés à leur domaine d'intérêt, en synthétisant et en évaluant des informations et arguments issus de plusieurs sources. Ils ont un bon éventail de vocabulaire pour les sujets liés à leur domaine ainsi que pour la plupart des thèmes généraux.

C1 – Les apprenants de ce niveau peuvent produire des textes clairs, bien structurés sur des sujets complexes, en soulignant les questions essentielles, en développant et en appuyant leurs points de vue de manière détaillée avec des arguments secondaires, des raisons et des exemples pertinents, et en concluant de manière appropriée. Ils savent aussi utiliser la structure et les conventions de divers genres, en adaptant le ton, le style et le registre selon le destinataire, le type de texte et le thème.

C2 – Les apprenants de ce niveau peuvent produire des textes clairs, fluides et complexes dans un style approprié et efficace, avec une structure logique qui aide le lecteur à identifier les points importants. Ils ont une excellente maîtrise d'un très large éventail lexical incluant des expressions idiomatiques et des tournures familières, et font preuve de sensibilité aux niveaux connotatifs de signification.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for written production in Estonian (ET)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Estonian** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 – Selle taseme õppijad suudavad anda teavet isiklikult olulistel teemadel (nt meeldimised ja mittemeeldimised, perekond, lemmikloomad), kasutades lihtsaid sõnu/viipeid ja põhilisi väljendeid. Õppijad suudavad moodustada ka lihtsaid üksikuid fraase ja lauseid.

A2 – Selle taseme õppijad suudavad toota lihtsate fraaside ja lausete jada, mis on seotud lihtsate sidesõnadega nagu „ja“, „aga“ ja „sest“. Neil on piisav sõnavara põhiliste suhtlusvajaduste ja lihtsate ellujäämisvajaduste rahuldamiseks.

B1 – Selle taseme õppijad suudavad koostada arusaadavaid, seotud tekste tuttavatel teemadel oma huvivaldkonnas, sidudes lühemaid üksikuid elemente lineaarseks järjestuseks. Neil on hea sõnavara tuttavate teemade ja igapäevaste olukordade kirjeldamiseks.

B2 – Selle taseme õppijad suudavad koostada selgeid ja üksikasjalikke tekste erinevatel nende huvivaldkonnaga seotud teemadel, sünteesides ja hinnates teavet ja argumente mitmest allikast. Neil on hea sõnavara oma valdkonnaga seotud teemadeks ning enamike üldiste teemade jaoks.

C1 – Selle taseme õppijad suudavad koostada selgeid ja hästi struktureeritud tekste keerukatel teemadel, tuues esile olulised küsimused, laiendades ja toetades seisukohti üksikasjalikult koos täiendavate punktide, põhjuste ja asjakohaste näidetega ning lõpetades sobiva järeldusega. Samuti suudavad nad kasutada erinevate žanrite struktuuri ja konventsioone ning varieerida tooni, stiili ja registrit vastavalt adressaadile, tekstiliigile ja teemale.

C2 – Selle taseme õppijad suudavad koostada selgeid, sujuvaid ja keerukaid tekste sobivas ja tõhusas stiilis ning loogilises struktuuris, mis aitab lugejal tuvastada olulisi punkte. Neil on väga lai sõnavara, mis sisaldab idioome ja kõnekeelseid väljendeid; nad tunnetavad ka tähenduse konnotatiivseid tasandeid.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for written production in Portuguese (PT)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Portuguese** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 – Os aprendentes deste nível conseguem fornecer informações sobre assuntos de relevância pessoal (por exemplo, gostos e preferências, família, animais de estimação) usando palavras/sinais simples e expressões básicas. Também conseguem produzir frases e expressões simples e isoladas.

A2 – Os aprendentes deste nível conseguem produzir uma série de frases e expressões simples ligadas por conectores básicos como "e", "mas" e "porque". Têm vocabulário suficiente para expressar necessidades comunicativas básicas e lidar com necessidades simples de sobrevivência.

B1 – Os aprendentes deste nível conseguem produzir textos simples e coerentes sobre uma variedade de temas familiares dentro de seu campo de interesse, ligando uma série de elementos mais curtos em sequência linear. Possuem um bom repertório de vocabulário relacionado a temas familiares e situações do cotidiano.

B2 – Os aprendentes deste nível conseguem produzir textos claros e detalhados sobre uma variedade de assuntos relacionados ao seu campo de interesse, sintetizando e avaliando informações e argumentos de várias fontes. Têm um bom vocabulário para assuntos relacionados à sua área e à maioria dos temas gerais.

C1 – Os aprendentes deste nível conseguem produzir textos claros e bem estruturados sobre temas complexos, destacando os pontos relevantes, desenvolvendo e sustentando pontos de vista com argumentos secundários, razões e exemplos pertinentes, e encerrando com uma conclusão apropriada. Também conseguem empregar a estrutura e as convenções de diferentes gêneros textuais, variando o tom, o estilo e o registro conforme o destinatário, o tipo de texto e o tema.

C2 – Os aprendentes deste nível conseguem produzir textos claros, fluidos e complexos em um estilo apropriado e eficaz, com uma estrutura lógica que ajuda o leitor a identificar os pontos significativos. Têm um excelente domínio de um repertório lexical muito amplo, incluindo expressões idiomáticas e coloquialismos, e demonstram consciência dos níveis conotativos de significado.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

## CEFR specifications for written production in Arabic (ar)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Arabic** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 - يمكن للمتعلمين في هذا المستوى تقديم معلومات حول مواضيع ذات صلة شخصية (مثل: ما يحبونه وما يكرهونه، العائلة، الحيوانات الأليفة) باستخدام كلمات/إشارات بسيطة وتعبيرات أساسية. كما يمكنهم أيضًا إنتاج عبارات وجمل بسيطة ومفردة.

A2 - يمكن للمتعلمين في هذا المستوى إنتاج سلسلة من العبارات والجمل البسيطة المترابطة باستخدام أدوات ربط بسيطة مثل "و"، "لكن"، و"لأن". لديهم مفردات كافية للتعبير عن الاحتياجات التواصلية الأساسية والتعامل مع احتياجات البقاء البسيطة.

B1 - يمكن للمتعلمين في هذا المستوى إنتاج نصوص مترابطة وبسيطة حول مجموعة من المواضيع المألوفة في مجال اهتمامهم، عن طريق ربط سلسلة من العناصر المنفصلة الأقصر في تسلسل خطي. لديهم مجموعة جيدة من المفردات المتعلقة بالمواضيع المألوفة والمواقف اليومية.

B2 - يمكن للمتعلمين في هذا المستوى إنتاج نصوص واضحة ومفصلة حول مجموعة متنوعة من المواضيع المرتبطة بمجال اهتمامهم، مع تلخيص وتقييم المعلومات والحجج من عدد من المصادر. لديهم مجموعة جيدة من المفردات المتعلقة بالأمور المرتبطة بمجالهم ومعظم المواضيع العامة.

C1 - يمكن للمتعلمين في هذا المستوى إنتاج نصوص واضحة ومبنية بشكل جيد حول مواضيع معقدة، مع إبراز القضايا البارزة ذات الصلة، وتوسيع النقاط المدعومة بأفكار فرعية وأسباب وأمثلة مناسبة، والانتهاء بخاتمة مناسبة. كما يمكنهم استخدام بنية واتفاقيات مجموعة متنوعة من الأنماط النصية، وتغيير النغمة والأسلوب والسجل حسب المتلقي ونوع النص والموضوع.

C2 - يمكن للمتعلمين في هذا المستوى إنتاج نصوص واضحة، وسلسة، ومعقدة بأسلوب مناسب وفعال وبنية منطقية تساعد القارئ على تحديد النقاط المهمة. لديهم تحكم جيد في مجموعة واسعة جدًا من المفردات تشمل التعابير الاصطلاحية والكلمات العامية؛ ويُظهرون وعيًا بمستويات المعنى الضمنية.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

**CEFR specifications for written production in Hindi (HI)**

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Hindi** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 - इस स्तर के शिक्षार्थी व्यक्तिगत महत्व के विषयों (जैसे पसंद और नापसंद, परिवार, पालतू जानवर) के बारे में सरल शब्दों/संकेतों और बुनियादी अभिव्यक्तियों का उपयोग करके जानकारी दे सकते हैं। ये शिक्षार्थी सरल, अलग–अलग वाक्य और वाक्यांश भी बना सकते हैं।

A2 - इस स्तर के शिक्षार्थी सरल वाक्यांशों और वाक्यों की एक श्रृंखला बना सकते हैं, जो “और”, “लेकिन” और “क्योंकि” जैसे सरल संयोजको द्वारा जुड़े होते हैं। इनके पास बुनियादी संप्रेषण आवश्यकताओं और सरल जीवन आवश्यकताओं को पूरा करने के लिए पर्याप्त शब्दावली होती है।

B1 - इस स्तर के शिक्षार्थी परिचित विषयों पर, जो उनके रुचि के क्षेत्र में आते हैं, सीधे और जुड़े हुए पाठ बना सकते हैं, जिसमें छोटे–छोटे अलग–अलग तत्वों को एक रैखिक क्रम में जोड़ा गया होता है। इनके पास परिचित विषयों और रोज़मर्रा की परिस्थितियों से संबंधित शब्दों का अच्छा भंडार होता है।

B2 - इस स्तर के शिक्षार्थी अपने रुचि क्षेत्र से संबंधित विभिन्न विषयों पर स्पष्ट, विस्तृत पाठ बना सकते हैं, जिसमें कई स्रोतों से जानकारी और तर्को को संयोजित और मूल्यांकित किया जाता है। इनके पास अपने क्षेत्र से संबंधित मामलों और अधिकांश सामान्य विषयों के लिए अच्छी शब्दावली होती है।

C1 - इस स्तर के शिक्षार्थी जटिल विषयों पर स्पष्ट, अच्छी तरह से संरचित पाठ बना सकते हैं, जिनमें महत्वपूर्ण मुद्दों को रेखांकित किया जाता है, और विचारों को सहायक बिंदुओं, कारणों और उपयुक्त उदाहरणों के साथ विस्तारपूर्वक प्रस्तुत किया जाता है, और अंत में एक उपयुक्त निष्कर्ष दिया जाता है। वे विभिन्न शैलियों की संरचना और परंपराओं का प्रयोग भी कर सकते हैं, और श्रोता, पाठ के प्रकार और विषय के अनुसार शैली, स्वर और औपचारिकता को समायोजित कर सकते हैं।

C2 - इस स्तर के शिक्षार्थी स्पष्ट, सहज प्रवाह वाले और जटिल पाठ बना सकते हैं जो उपयुक्त और प्रभावी शैली में होते हैं और जिनकी तार्किक संरचना पाठक को मुख्य बिंदुओं की पहचान करने में सहायता करती है। इनके पास बहुत विस्तृत शब्दावली का अच्छा नियंत्रण होता है, जिसमें मुहावरे और बो–लचाल की अभिव्यक्तियाँ शामिल होती हैं; वे अर्थ की व्यंजक (connotative) परतों के प्रति भी सजग होते हैं।

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for written production in Russian (RU)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Russian** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 — Обучающиеся на этом уровне могут сообщать информацию на личные темы (например, о своих предпочтениях, семье, домашних животных), используя простые слова/жесты и базовые выражения. Также они могут составлять простые отдельные фразы и предложения.

A2 — Обучающиеся на этом уровне могут строить серию простых фраз и предложений, соединённых с помощью простых союзов, таких как «и», «но» и «потому что». У них есть достаточный словарный запас для выражения базовых коммуникативных потребностей и для решения простых бытовых задач.

B1 — Обучающиеся на этом уровне могут создавать понятные связные тексты на знакомые темы в рамках своей области интересов, объединяя серию коротких, отдельных элементов в линейную последовательность. У них хороший запас слов, связанных с повседневными ситуациями и знакомыми темами.

B2 — Обучающиеся на этом уровне могут писать чёткие и подробные тексты по различным темам, связанным с их сферой интересов, обобщая и оценивая информацию и аргументы из нескольких источников. У них хороший словарный запас по тематике своей области и большинству общих тем.

C1 — Обучающиеся на этом уровне могут создавать чёткие, хорошо структурированные тексты на сложные темы, подчёркивая важные аспекты, развивая и обосновывая свою точку зрения с помощью дополнительных аргументов, причин и релевантных примеров, и завершать текст уместным заключением. Они также могут применять структуру и нормы различных жанров, варьируя тон, стиль и регистр в зависимости от адресата, типа текста и темы.

C2 - Обучающиеся на этом уровне могут создавать чёткие, плавные и сложные тексты в уместном и эффективном стиле, с логичной структурой, которая помогает читателю выделять важные моменты. У них отличное владение очень широким лексическим запасом, включая идиоматические выражения и разговорную лексику; они осознают коннотативные уровни значений.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer:

### CEFR specifications for written production in Welsh (CY)

You are an expert in language proficiency classification based on the Common European Framework of Reference for Languages (CEFR). Your task is to analyze the given **Welsh** text or narrative and determine the best CEFR level [A1, A2, B1, B2, C1, or C2] based on the CEFR descriptors of reading comprehension of learners below:

A1 – Gall dysgwyr ar y lefel hon roi gwybodaeth am faterion o berthnasedd personol (e.e. pethau maen nhw'n eu hoffi a'u casáu, teulu, anifeiliaid anwes) gan ddefnyddio geiriau/arwyddion syml ac ymadroddion sylfaenol. Gall dysgwyr hefyd gynhyrchu brawddegau ac ymadroddion syml, arwahanol.

A2 – Gall dysgwyr ar y lefel hon gynhyrchu cyfres o ymadroddion a brawddegau syml wedi'u cysylltu gan gysyllteiriau syml fel "a", "ond" a "oherwydd". Mae gan ddysgwyr eirfa ddigonol i fynegi anghenion cyfathrebu sylfaenol ac i ymdopi ag anghenion goroesi syml.

B1 – Gall dysgwyr ar y lefel hon gynhyrchu testunau cysylltiedig, uniongyrchol ar ystod o bynciau cyfarwydd o fewn eu maes diddordeb, drwy gysylltu cyfres o elfennau byrrach ar wahân i mewn i ddilyniannol linol. Mae ganddynt ystod dda o eirfa sy'n ymwneud â phethau cyfarwydd a sefyllfaoedd bob dydd.

B2 – Gall dysgwyr ar y lefel hon gynhyrchu testunau clir, manwl ar amrywiaeth o bynciau sy'n gysylltiedig â'u maes diddordeb, gan gyfuno a gwerthuso gwybodaeth a dadleuon o sawl ffynhonnell. Mae ganddynt ystod dda o eirfa ar gyfer materion sy'n gysylltiedig â'u maes ac ar gyfer y rhan fwyaf o bynciau cyffredinol.

C1 – Gall dysgwyr ar y lefel hon gynhyrchu testunau clir, wedi'u strwythuro'n dda ar bynciau cymhleth, gan amlygu'r materion perthnasol, ehangu a chefnogi safbwyntiau'n fanwl gyda phwyntiau ategol, rhesymau ac enghreifftiau perthnasol, a gorffen gyda chasgliad priodol. Gallant hefyd ddefnyddio strwythur a chonfensiynau amrywiaeth o genres, gan amrywio'r naws, arddull a chofrestr yn ôl y derbynnydd, math y testun a'r thema.

C2 – Gall dysgwyr ar y lefel hon gynhyrchu testunau clir, esmwyth a chymhleth mewn arddull briodol ac effeithiol ac mewn strwythur resymegol sy'n helpu'r darllenydd i nodi pwyntiau arwyddocaol. Mae ganddynt reolaeth dda dros eirfa eang iawn gan gynnwys ymadroddion idiomatig a llafariad; maent yn dangos ymwybyddiaeth o lefelau ystyron cynhennus.

Provide only the CEFR level as output directly, without explanation or justification.

Text: «TEXT»

Answer: