# UNIVERSALCEFR: Enabling Open Multilingual Research on Language Proficiency Assessment

**Joseph Marvin Imperial**[1,3], **Abdullah Barayan**[2,14], **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, [14]King Abdulaziz University

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

## Abstract

We introduce **UNIVERSALCEFR**, a large-scale multilingual and multidimensional dataset of texts annotated with CEFR (Common European Framework of Reference) levels in 13 languages. To enable open research in 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 modelling across tasks and languages. To demonstrate its utility, we conduct benchmarking 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 support using linguistic features and fine-tuning pretrained models in multilingual CEFR level assessment. Overall, UNIVERSALCEFR aims to establish best practices in data distribution for 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

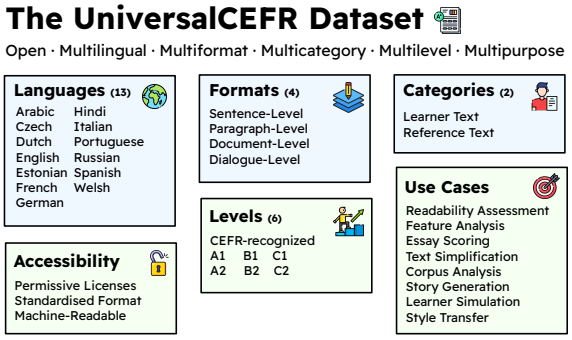


Figure 1: Overview of the contributions of the **UNIVERSALCEFR** dataset, highlighting its **diverse structural coverage**—spanning language, format, category, and CEFR level—as well as its **accessibility and interoperability** for downstream tasks and use cases enabled by permissive licenses and standardized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

## 1 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] devel-

[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† | 6 | text | unrestricted | no | none |
| Corpora @ UCLouvain | 31† | 9 | text, audio, video | request per corpus | no | yes |
| CLARIN L2 Learner Corpora | 75† | 34 | text, video | request per corpus | no | yes |
| Learner Language (Språkbanken) | 15† | 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. † indicates that only a 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.

oped by the Council of Europe. CEFR offers a 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-resource 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.

**CEFR Assessment and Standardization.** 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. Likewise, in data collection standardization, CLARIN released the Core Metadata Schema for Learner Corpora (LC-meta), which aims to provide a structured method with a specific emphasis on capturing metadata of collected learner texts, focusing on learner background, context, and individual differences (Paquot et al., 2024).

## 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. Unifying diverse resources under a consistent format enables reproducible and scalable research across linguistics, NLP, and education. In this section, we outline the dataset's design principles, detail the data collection and standardization pipeline, provide key statistics, and present a linguistic feature analysis that supports downstream modelling.

[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/