A Comparative Study of Feature Representations for Spanish L2 Text Complexity Across Pedagogical and Authentic Corpora

Un estudio comparativo de las representaciones de características para la complejidad textual del español como L2 en corpus pedagógicos y auténticos

Daniel Mora,[1] **María José Domínguez Vázquez,**[2] **Iván Arias-Arias,**[2] **Idalete Dias,**[3] **Vítor Míguez-Rego,**[2] **Carlos Valcárcel Riveiro**[4]
[1]Leibniz Institute for Science and Mathematics Education, Germany
[2]Instituto da Lingua Galega, Universidade de Santiago de Compostela, Spain
[3]Centro de Estudos Humanísticos, Universidade do Minho, Portugal
[4]Instituto de investigación Lingua, University of Vigo, Spain
mora@leibniz-ipn.de, {majo.dominguez,ivanarias.arias}@usc.es
idalete@elach.uminho.pt, vitor.miguez@usc.gal, carlos.valcarcel@uvigo.gal

Abstract: This study investigates the effectiveness of three feature representations for Spanish L2 CEFR level prediction: traditional readability metrics, CEFR-based lexical overlap features, and unigram TF-IDF. We evaluate these features individually and in combination under varying data conditions across two types of datasets: an authentic learner corpus (CAES) and two pedagogical datasets (HablaCultura and Kwiziq). Results show that distributional features achieve the highest performance when sufficient training data is available. In contrast, in smaller or cross-dataset settings, traditional readability metrics combined with CEFR-based lexical features provide more stable performance. We also observe limited generalization across dataset types, particularly when transferring from learner-produced to pedagogical texts. Overall, the findings highlight the need to adapt feature representations to data conditions and domain characteristics for Spanish L2 CEFR prediction.
Keywords: text readability assessment, Spanish L2, learner-produced texts, pedagogical texts.

Resumen: Este estudio investiga la efectividad de tres representaciones de características para la predicción de niveles CEFR en español como L2: métricas tradicionales de legibilidad, características léxicas basadas en solapamiento alineado con categorías CEFR y TF-IDF de unigramas. Evaluamos estas características de forma individual y en combinación bajo distintas condiciones en dos tipos de corpora: un corpus auténtico de aprendices (CAES) y dos corpora pedagógicos (HablaCultura y Kwiziq). Los resultados muestran que las características distribucionales alcanzan el mejor rendimiento cuando se dispone de suficientes datos de entrenamiento. En cambio, con pocos datos o entre corpora, las métricas de legibilidad combinadas con el solapamiento léxico basado en listas CEFR resultan más estables. Hay también una generalización limitada entre tipos de corpora al transferir textos producidos por aprendices a textos pedagógicos. Los resultados destacan la necesidad de adaptar las representaciones de características a las condiciones de los datos y a las características del corpus para la predicción de niveles CEFR en español como L2.
Palabras clave: evaluación de la legibilidad textual, español como L2, textos producidos por aprendices, textos pedagógicos.

1 Introduction

Analyzing text complexity and readability serves clear real-world purposes with direct implications for educational practice (Zalmout, Saddiki, and Habash, 2016). The ability to automatically assess text difficulty

is essential for supporting pedagogical and social goals, such as improving adult reading skills, which is the aim of contemporary European projects like iRead4Skills (Reis et al., 2024; Aissa et al., 2025; Rey and Garcia, 2025). This need has driven the development of automatic text classification tools (Rodríguez Rey, 2024) and research into the specific nature of textual complexity in foreign language (L2) teaching contexts (François et al., 2014). For language teachers, material designers, and L2 learners, selecting texts that are appropriately challenging, neither prohibitively difficult nor insufficiently demanding, remains a critical endeavour.

However, automatic text difficulty assessment remains challenging. While traditional readability formulas like Flesch–Kincaid rely on surface features (e.g., sentence length, number of syllables), they often yield unreliable results, especially in L2 contexts, as they fail to capture linguistic factors that increase difficulty from the learner's perspective (Tanprasert and Kauchak, 2021).

Regarding L2 contexts, we distinguish between two types of texts depending on the perspective: learning or teaching. Authentic or learner-produced texts are written by learners, thus come with high language variation, errors, and tend to fall below the assigned level of competence. Pedagogical texts or learner-oriented texts, on the contrary, have controlled vocabulary repetition, structural simplification, and tend to be calibrated slightly above the assigned level as they are designed as challenging input. This introduces a systematic asymmetry between both corpus types that goes beyond mere domain or register differences.

This study investigates the effectiveness of different feature representations for predicting Spanish L2 text readability according to CEFR levels on pedagogical and authentic corpora. In particular, we compare three complementary feature representations: (i) traditional readability metrics capturing surface-level text complexity, (ii) CEFR-aligned lexical features derived from expert-curated resources, and (iii) data-driven distributional features based on unigram TF-IDF.

The evaluation is conducted on three CEFR-annotated Spanish L2 corpora: *Hablacultura* and *Kwiziq* (Vásquez-Rodríguez et al., 2022), which are designed for Spanish language instruction, and the *Corpus de Aprendices de Español* (Rojo and Martínez, 2016), a large-scale learner corpus. This setup allows us to examine performance across different text types, namely pedagogically curated materials and authentic learner productions. Given the diverse nature of these datasets, we also account for cross-domain generalizability and the effect of dataset size in the models' performance.[1]

Considering CEFR level prediction as a multi-class classification task, we systematically assess the contribution of each feature group by comparing their individual performance as well as their combinations through controlled ablation experiments. In addition to within-dataset evaluation, we analyse the robustness of feature representations under varying data conditions. Specifically, we test their ability to generalize across types of texts by training on learner-produced texts and evaluating on pedagogical corpora. We also control for differences in dataset size through downsampling, allowing us to isolate the impact of training data availability from the effectiveness of the feature representations.

The main research questions addressed in this study are, thus:

- RQ1: How does combining traditional readability metrics, CEFR-based lexical features, and distributional representations affect Spanish L2 CEFR level classification performance compared to individual feature groups?
- RQ2: To what extent do feature representations generalize across pedagogical and authentic texts, and how is their effectiveness affected by variations in dataset size?

In what follows, Section 2 reviews related work. Section 3 describes the datasets, text features, and experimental setup. Section 4 presents the results, and Section 5 concludes with a discussion of the findings.

2 Background

Analyzing text complexity and readability serves critical pedagogical and social purposes: improving adult literacy (Monteiro et

[1]Source code available at https://github.com/GrupoPortlex/ReadabilityESMAS

al., 2023; Reis et al., 2024), enhancing patient comprehension of medical texts (Scholz and Wenzel, 2025), enabling automatic text classification (Rodríguez Rey, 2024; Rey and Garcia, 2025), and informing foreign language teaching practices (François et al., 2014). These applications require quantitative and qualitative assessment parameters, annotated datasets (Ribeiro, Mamede, and Baptista, 2024), and specialized tools such as MultiAzterTest (Bengoetxea and Gonzalez-Dios, 2021).

The representation of text remains a critical issue in supervised CEFR prediction, as datasets vary substantially in size, languages differ in their linguistic properties, and text difficulty depends on interacting linguistic, extralinguistic, and reader-related factors. This variability makes it challenging to design feature representations that both generalize across settings and capture the multidimensional nature of readability.

2.1 Traditional Readability Metrics

Traditional readability metrics rely on length- and frequency-based features. Early work established these variables as proxies for text difficulty (Thorndike, 1921; Lively and Pressey, 1923; Vogel and Washburne, 1928), and led to widely used formulas such as Flesch Reading Ease (Flesch, 1948), Flesch-Kincaid, Gunning Fog, Dale–Chall (Dale and Chall, 1948), and SMOG (McLaughlin, 1969). Their main strength lies in their simplicity, interpretability, and computational efficiency, which makes them strong and widely used baselines (Bestgen, 2020; Martinc, Pollak, and Robnik-Šikonja, 2021; Vajjala, 2022).

However, most formulas were developed for English and native speakers, and they largely overlook semantic, lexical, and pragmatic aspects of language (Bengoetxea and Gonzalez-Dios, 2021). As a result, they fail to capture the multifaceted nature of text complexity, particularly in L2 contexts (Zhou, Jeong, and Green, 2017). Cross-linguistic studies (Azpiazu and Pera, 2019) and multilingual initiatives further show that sentence and word length alone are insufficient, motivating the inclusion of richer dimensions such as lexical transparency, conceptual familiarity, domain specialization, abstraction level, and technical density (Hiebert, 2011; François et al., 2014; Monteiro et al., 2023).

2.2 The CEFR Framework in Language Pedagogy

The Common European Framework of Reference for Languages (CEFR) establishes six proficiency levels (A1–C2) defined through "Can-Do" descriptors of learner capabilities that are widely adopted for curriculum design, assessment, and materials development across languages and educational contexts. This levelling system is influential in vocabulary instruction given the correlation between lexical knowledge and CEFR proficiency (Milton and Alexiou, 2009; Milton, 2010; Coniam, Milanovic, and Zhao, 2022; Lew and Wolfer, 2024).

Researchers have developed CEFR-aligned assessment tools incorporating lexical profiles and grammatical complexity measures more relevant to language learners' actual reading experiences (Xia, Kochmar, and Briscoe, 2016; Arase, Uchida, and Kajiwara, 2022; Bestgen, 2020). For instance, the English Vocabulary Profile (Capel, 2015) and FLElex for French (François et al., 2014) demonstrate growing sophistication in CEFR-based text analysis across linguistic contexts that are relevant to multilingual projects like ESMAS-ES^{+} (Domínguez Vázquez, 2025), where consistent levelling across Spanish, French, German, and Galician requires robust theoretical foundations.

Curated vocabulary lists play a crucial role in operationalizing lexical complexity for CEFR-based assessment. Unlike purely frequency-based resources, which reflect general native language use, curated lists are constructed through a combination of corpus evidence, expert pedagogical judgment, and explicit alignment with CEFR descriptors, ensuring that the vocabulary they contain reflects the expected acquisition trajectory of L2 learners at each proficiency stage.

From a practical standpoint, these lists serve a dual purpose. In language teaching, they inform material selection, syllabus design, and vocabulary instruction by providing level-appropriate targets. In automatic text assessment, they enable the operationalization of CEFR levels as measurable lexical features, as the degree of overlap between a text's vocabulary and a given level's list pro-

vides a direct and interpretable signal of its alignment with that proficiency stage.

2.3 Automatic CEFR Assessment

The automatic assessment of text complexity is usually considered a classification task in which one of the CEFR proficiency levels is assigned to a text. It poses several challenges, including the scarcity of annotated data, variability across domains and learner populations, and the need to adequately derive feature representations from text that reflect the actual complexity of the text in a meaningful way. Existing approaches can be broadly grouped into three paradigms: (i) feature-based models relying on handcrafted linguistic indicators, (ii) neural models based on distributed representations, and (iii) generative large language models.

Feature-based approaches Early work in CEFR prediction predominantly relied on surface linguistic features. These include traditional readability metrics, lexical diversity, and morphosyntactic indicators. Vajjala and Rama (2018) explored monolingual, multilingual, and cross-lingual CEFR classification, showing that feature-based models can generalize across languages to some extent, suggesting the presence of language-independent signals of proficiency. Bestgen (2020) demonstrated that basic indicators such as text length and readability formulas remain strong predictors of CEFR levels, particularly in monolingual settings.

Neural approaches More recent research has shifted toward neural architectures, particularly transformer-based models that leverage contextualized embeddings. Schmalz and Brutti (2021) showed that such models achieve strong performance on CEFR classification tasks, benefiting from their ability to encode syntactic and semantic information in a unified representation. Beyond document-level classification, Arase, Uchida, and Kajiwara (2022) proposed sentence-level CEFR annotation and demonstrated that proficiency distinctions can be reliably detected at this level, enabling more precise modeling of learner competence. Recently, Rodríguez Rey and Garcia (2025) show that pre-trained models offer competitive performance compared to traditional features on the Galician language.

Generative large language models Recent work has investigated the use of large language models (LLMs) for CEFR prediction. Yancey et al. (2023) showed that models such as GPT-4 can approximate human CEFR judgments, indicating that proficiency-related knowledge is implicitly encoded in large-scale pretrained models. He and Li (2024) proposed a zero-shot framework for CEFR prediction, learning language-agnostic representations that generalize across multiple languages. However, these approaches raise concerns regarding robustness, consistency, and fairness, particularly across different learner populations and text types (Benedetto et al., 2025; Imperial et al., 2025).

CEFR assessment for Spanish In contrast to English, CEFR prediction for Spanish remains comparatively underexplored. Existing studies have primarily focused on feature-based approaches applied to learner corpora. del Río (2019) show that morphosyntactic features, especially part-of-speech distributions, are highly informative for proficiency classification. Similarly, Miaschi et al. (2020) modelled the development of written competence in L2 Spanish learners, demonstrating that automatically extracted linguistic features align well with pedagogical notions of progression.

Due to the limited availability of CEFR-annotated Spanish data, Vásquez-Rodríguez et al. (2022) introduced a large-scale benchmark for Spanish readability assessment, including texts targeted at L2 learners. However, it remains underexplored how different feature sets perform across varying text types, such as authentic and pedagogically adapted texts.

3 Methodology

We investigate the relevance of different group of features for automatic CEFR assessment on Spanish L2 datasets. The group of features are traditional readability metrics, CEFR-aligned lexical features derived from resources annotated by experts, and data-driven distributional features based on TF-IDF.

The task is formulated as a multi-class classification problem. We use Quadratic Weighted Kappa (QWK) (Cohen, 1968) as our primary evaluation metric, since it accounts for the ordinal nature of the labels and

penalizes misclassifications proportionally to their distance. For comparability with prior work, we also report different metrics on Appendix A.

The main criteria when selecting a model were its interpretability, the required amount of training data, and the ease of reproducibility across different domains. We compare several baselines and chose the best hyperparameters based on preliminary experiments (3-fold stratified cross-validation) using ngram TF-IDF features. A Logistic Regression model with L2 regularization[2] is adopted based on overall performance and stability. The following subsections describe the datasets used, how features were derived, and experimental design.

3.1 Datasets

We conduct our experiments on three complementary corpora of Spanish texts compiled by Vásquez-Rodríguez et al. (2022). These datasets include both pedagogically curated materials and authentic learner-produced texts, providing variation in writing style, readability, and sample size. Our goal in using different corpora is to examine how various text representations model authentic texts with uncontrolled complexity and errors, in comparison to texts that are deliberately simplified or that gloss over such complexity.

The datasets are collected using the Hugging Face `datasets`[3] library (Lhoest et al., 2021). Duplicate instances and samples with missing text or CEFR labels are removed. For features that depend on lexical normalization, texts are lemmatized using the `stanza` library (Qi et al., 2020) to reduce morphological variation. Other features are computed from the original text.

Corpus de Aprendices de Español (CAES) An authentic corpus of learner Spanish created by Rojo and Martínez (2016). It consists of 20,636 texts written by students of Spanish as a foreign language and annotated with CEFR category (A1–C1). Because CAES reflects language produced by L2 learners, it contains authentic errors and varying styles, with overall level corresponding to what a learner at that CEFR stage can produce. This corpus provides insight into the lexical and syntactic choices learners make at each level in unguided writing, which differs from the other two datasets that are meant to be used for teaching purposes.

HablaCultura A pedagogical text corpus consisting of reading materials from an educational website for Spanish learners (hablacultura.com) annotated with CEFR category (A2–C1). These texts cover cultural and general topics (e.g., arts, customs, science) intended for learners. Because they are prepared for teaching, they typically use controlled language: even at higher levels, texts are designed to be comprehensible to learners, reinforcing learned vocabulary from preceding levels.

Kwiziq A pedagogical dataset from the online platform Kwiziq (spanish.kwiziq.com), which provides lessons and readings for Spanish learners annotated with CEFR level (A1–C1). The Kwiziq corpus represents short reading comprehension passages and grammar explanations, carefully written to include level-appropriate vocabulary and syntactic structures.

Table 1 shows the distribution of CEFR labels across datasets. Overall, the Hablacultura and Kwiziq datasets are skewed toward the extreme categories, whereas CAES is skewed toward the lower proficiency levels. There is also a substantial difference in dataset size, as well as class imbalance within each dataset.

3.2 Text Features

We derive three group of features that capture complementary aspects of linguistic complexity and lexical proficiency: traditional readability metrics, CEFR-aligned lexical features, and distributional representations based on TF-IDF.

3.2.1 Readability Metrics

We compute four readability metrics using the `textstat` library:[4] Fernández–Huerta, Szigriszt–Pazos, Gutiérrez–Polini, and Crawford. Additionally, we add basic text statistics, including syllable count, lexicon count, sentence count, words per sentence, average syllables per word, and the number of difficult words, which accounts for a total of 10 features.

[2] `solver="lbfgs", C=1.0`

[3] `lmvasque/hablacultura`, `lmvasque/kwiziq`, `UniversalCEFR/caes_es`

[4] https://github.com/textstat/textstat

Dataset	A1	A2	B1	B2	C1	Total
CAES	5553 (27%)	5708 (28%)	4489 (22%)	3100 (15%)	1786 (9%)	20636
HablaCultura	0 (0%)	33 (5%)	245 (38%)	273 (42%)	99 (15%)	650
Kwiziq	21 (10%)	33 (16%)	54 (26%)	59 (29%)	39 (19%)	206

Table 1: Distribution of CEFR levels across datasets.

These features are computed directly on the original texts and capture surface-level linguistic properties, such as lexical density, sentence length, and syllabic complexity. As such, they provide an interpretable representation of textual difficulty, without relying on deeper syntactic or semantic analysis.

3.2.2 CEFR-aligned lexical features

In addition to readability metrics, we also compute CEFR-aligned lexical features. This is measured by the overlap between the text and curated CEFR-based vocabulary lists. In our experiments, we use the lists provided by the Instituto Cervantes as part of its Curricular Plan,[5] which defines the vocabulary learners are expected to know by the end of each proficiency level and guides both textbook vocabulary selection and language examination content. We downloaded and manually checked the entries.

These lists include both single-word entries and multi-word expressions labelled with a CEFR label (A1–B2). Table 2 shows examples from each CEFR category along with their distribution.

For each CEFR level, all associated lemmas are combined into a single case-insensitive pattern that matches complete words, and multi-word expressions are treated as single lexical units to ensure full overlap when they occur. Overlap is defined as the count of lemmas from each CEFR-specific list that appear in the lemmatized text, resulting in one count per CEFR level. This yields four features per text, each representing the frequency of vocabulary corresponding to a given CEFR level, and thus quantifies how strongly the text aligns with different proficiency levels based on its lexical composition.

Even though CEFR-based lexical lists provide structured and curated information, their overlap with real language use in texts is relatively limited. In practice, the number of matched items per text remains low, typically below 20 on average across datasets per CEFR level, which is largely a consequence of the short length of the texts and the high lexical diversity of natural language.

A slight increase in overlap can be observed as CEFR level rises, reflecting the greater lexical richness of more advanced texts. However, this relationship is weak. One reason is that A1-level vocabulary is highly frequent and appears across all levels, introducing noise and reducing the discriminative power of overlap-based features. At the same time, higher-level vocabulary (e.g., B2) is comparatively sparse, so even in advanced texts only a small portion of the curated lists is actually observed. As a result, despite the curated lists containing several thousand lexical items, only a small subset is matched in any given text, limiting the effectiveness of simple overlap measures for distinguishing CEFR levels.

3.2.3 TF-IDF

Finally, distributional representations are obtained using TF-IDF (Ramos, 2003) unigram features computed using Scikit-Learn library (Pedregosa et al., 2011). These representations encode word usage patterns across the corpus and capture contextual information not directly reflected in lexical lists or readability metrics.

Conceptually, this feature group differs from CEFR-based curated list overlap in that it adopts a data-driven, bottom-up approach, rather than relying on predefined lexical inventories. It avoids the main drawback of curated lists, namely the Zipfian distribution of language use. However, it may introduce noise by reflecting corpus-specific patterns rather than generalizable linguistic properties.

The resulting high-weight n-grams (i.e., n-grams with high TF-IDF scores, representing terms that are frequent in a document yet discriminative across the corpus) can thus serve as empirical proxies for representative vocabulary, although they require post-processing to account for noise introduced by learner language, such as spelling errors or out-of-dictionary words.

[5] https://cvc.cervantes.es/ensenanza/

CEFR	Examples	Count	Percentage
A1	monumento, invierno, herramienta	485	11.61%
A2	sed, internet, gramo	575	13.76%
B1	conservador, carril, electrodoméstico	1182	28.28%
B2	hinduismo, papelera, representante	1940	46.35%
Total		4182	100%

Table 2: Examples of lexical items and distribution across CEFR levels according to the curated lists from Instituto Cervantes.

Table 3 presents the five highest-ranked unigrams for the CAES dataset. Even this reduced sample reveals clear differences across proficiency levels. At A1, the most salient items include basic, high-frequency vocabulary such as common nouns (*año*, *padre*) and frequent verbs (*gustar*). At intermediate levels, B1 already shows the presence of more structurally relevant items such as the conditional marker *si*. At higher levels, the vocabulary shifts towards more abstract and semantically complex terms, as illustrated by C1 items like *servicio* and *compañía*.

CEFR	Top unigrams
A1	año, gustar, tener, llamar, padre
A2	habitación, vacaciones, visitar, hotel, reservar
B1	equipaje, maleta, poder, si, pedir
B2	fumar, programa, público, lengua, universidad
C1	película, electricidad, servicio, factura, compañía

Table 3: Top 5 TF-IDF unigrams per CEFR level for the CAES dataset.

3.3 Experimental Design

To systematically assess the contribution and robustness of different feature groups under varying data conditions, we design three complementary experiments. First, we evaluate the predictive capacity of each feature group and their combinations within each dataset. Second, we analyze cross-dataset generalization by training on learner-produced texts and testing on pedagogical corpora. Finally, we control for dataset size by downsampling the largest corpus, allowing us to differentiate the effect of training data size from feature effectiveness.

Feature Group Comparison In the first experiment, we assess the predictive contribution of different feature groups evaluating on a 5-fold cross validation setup. We evaluate readability features, CEFR-based lexical overlap features, and TF-IDF unigram representations individually, as well as their combinations to examine potential complementarity. The goal is to identify which feature group, or combination of feature groups, provides the most effective representation for CEFR level prediction.

Cross-Dataset Generalization Motivated by the performance differences observed in Experiment 1 between learner-produced and pedagogically curated texts, we evaluate how the same feature group combinations are able to generalize across datasets with different characteristics. Specifically, we train the model on the CAES corpus and evaluate it on the pedagogical datasets HablaCultura and Kwiziq. This setup allows us to assess how well each feature group transfers from authentic learner-produced texts to pedagogically curated materials.

Training Size Effect Given the substantial size difference between CAES and the pedagogical datasets, we examine the impact of training data size on model performance by downsampling CAES and generating a learning curve. Sampling is performed in a stratified manner to preserve the original label distribution. At every training size, we test using 5-fold cross validation. The objective of this experiment is to determine whether performance differences across datasets are primarily driven by feature effectiveness or by the amount of available training data.

4 Results

This section presents the results in terms of QWK. For brevity, the feature groups compared are referred to as *readability*, *cefr-vocab*, and *TF-IDF*. Detailed results are depicted in Appendix A.

4.1 Feature Groups Comparison

Figure 1 presents the performance of each feature group across datasets, reported as

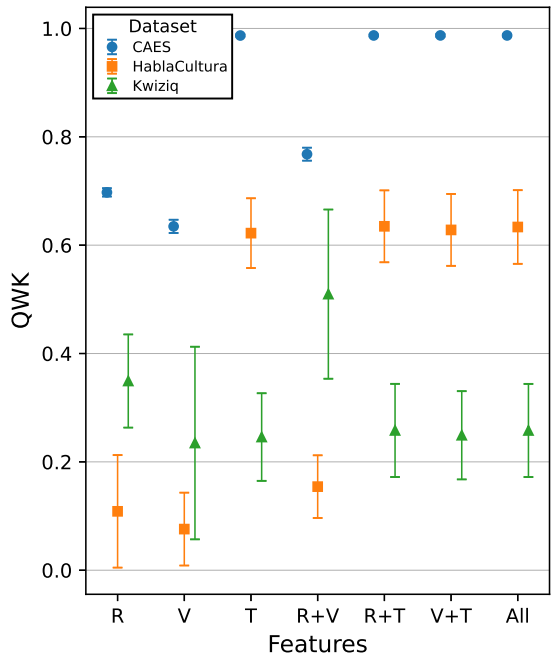

Figure 1: Performance (QWK) averaged over 5-fold cross-validation (mean and std) across feature configurations per dataset. R: readability, V: cefr-vocab, T: TF-IDF.

mean and standard deviation under a 5-fold cross-validation setup.

For the CAES corpus, TF-IDF features achieve the highest performance (0.98). Adding additional features (readability or cefr-vocab) yields only marginal improvements, indicating that n-gram representations alone capture most of the relevant signal in this setting. In contrast, removing n-grams leads to a drop in performance, from 0.98 to 0.69 (readability) and 0.63 (cefr-vocab), confirming their central role in this dataset.

This pattern is less consistent in the pedagogical datasets. In HablaCultura, n-gram features remain the strongest individual representation (0.62), with only slight improvements when combined with other features, suggesting limited complementarity. However, removing n-grams results in a substantial degradation in performance, dropping to 0.10 (readability) and 0.07 (cefr-vocab), indicating that surface-level metrics are insufficient to capture complexity, and that overlap-based features depend heavily on lexical coverage.

In Kwiziq, the advantage of n-grams is further reduced. TF-IDF features achieve relatively low performance (0.24), and combining them with other feature groups yields only minor improvements (e.g., 0.25 for readability+TF-IDF). The best performance is obtained with the combination of readability and cefr-vocab features (0.51), outperforming both readability (0.34) and cefr-vocab (0.23) when used individually. The confusion matrices in Appendix A show that their combination improves the classification of lower proficiency levels, whereas n-gram features tend to cluster predictions around the B1–B2 levels. The lower performance of n-gram-based representations suggests that distributional features are less effective on small datasets, which limits the reliability of learned word distributions.

4.2 Cross-Dataset Generalization

Considering the high performance of TF-IDF features on the CAES corpus, we evaluate their ability to generalize to pedagogical corpora. Figure 2 presents the comparison between cross-dataset performance (trained on CAES and tested on pedagogical corpus) and within-corpus baselines for each feature configuration (from Experiment 1). Overall, only TF-IDF features transfers effectively from CAES to Kwiziq, but not to HablaCultura, which suggests that there are distributional similarities that benefit the smaller dataset with a model trained on a larger dataset.

When testing on HablaCultura, cross-dataset performance is consistently low across all feature configurations. The best result is obtained with readability+TF-IDF (0.16), closely followed by all features (0.15) and TF-IDF (0.14). In contrast, the corresponding within-corpus results are substantially higher, with readability+TF-IDF reaching 0.63 and TF-IDF 0.62. This corresponds to performance drops of 0.47 and 0.48 points, respectively, highlighting the limited transferability of n-gram-based representations. As expected, feature configurations that already achieve low performance (readability and cefr-vocab) do not benefit from the cross-dataset setup.

When testing on Kwiziq, however, there is a different pattern. Only the feature configurations that include TF-IDF show improvement over their within-corpus baselines, with an average increase of 0.25. The configuration that includes all features shows the

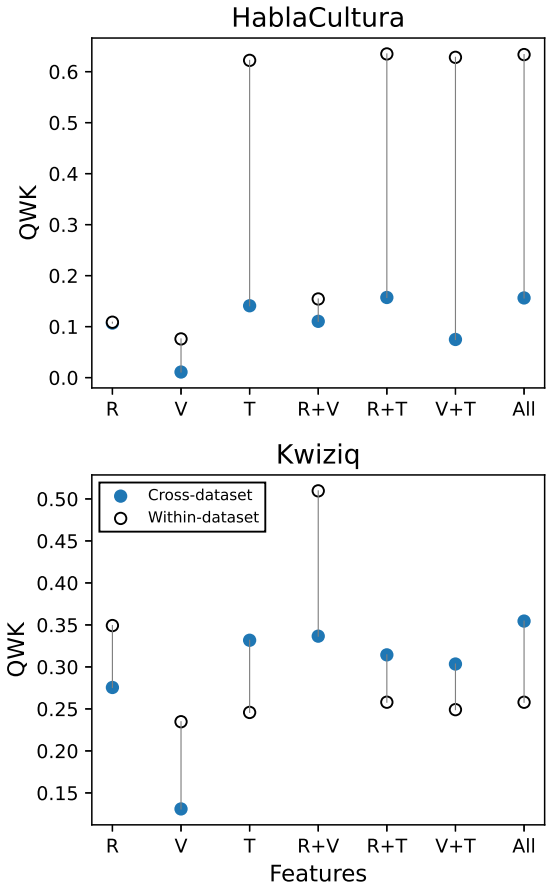

Figure 2: Performance (QWK) across feature configurations when trained on CAES and tested on HablaCultura or Kwiziq. Difference compared to within-corpus baseline from Experiment 1.

largest gain, rising from 0.25 to 0.35.

4.3 Training Size Effect

This experiment investigates whether the relative effectiveness of feature configurations depends on the amount of available training data, and to what extent. We combine two complementary analyses on the CAES corpus. First, we evaluate all feature group combinations under controlled training sizes by downsampling CAES to match the sizes of the pedagogical datasets. Second, we examine learning curves for two representative configurations across a range of training sizes.

Downsampling To simulate low-resource conditions, the CAES corpus is downsampled to 650 samples (matching HablaCultura) and 206 samples (matching Kwiziq). Figure 3 presents the resulting performance across all feature configurations.

Across both downsampled training sizes,

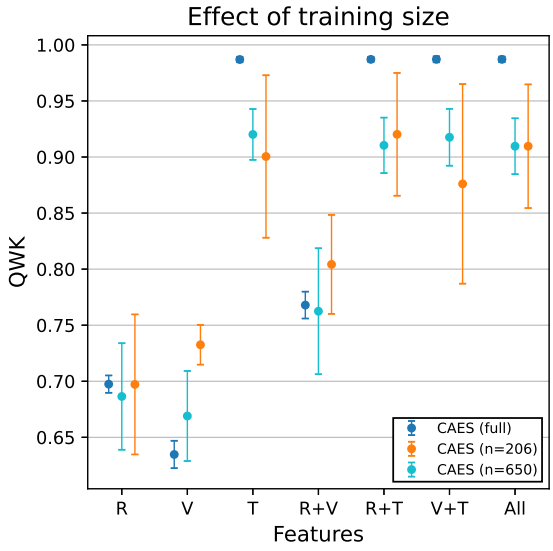

Figure 3: Performance (QWK) averaged over 5-fold cross-validation (mean and std) of all feature group combinations when CAES downsampled. R: readability, V: cefr-vocab, T: TF-IDF.

feature configurations that include TF-IDF achieve mean QWK scores above 0.85. The best result is obtained with TF-IDF alone (0.92). Increasing the amount of training data primarily reduces variability, leading to more stable results. The strong result despite the reduced training size also suggests that the texts belong to similar topics, indicating reduced variability, which in turn may affect the model's generalizability.

For readability metrics, adding more data does not improve mean performance (0.68), but only reduces variability. In contrast, for CEFR-based vocabulary features, smaller training sizes appear to yield better performance, with a difference of 7 points. However, given the differences observed across pedagogical corpora, cefr-vocab is highly sensitive to dataset-specific artifacts, possibly reflecting lexical matching effects rather than robust generalization.

Learning curve To further analyse how feature contributions change with increasing data, we examine learning curves for two representative configurations: TF-IDF and readability+cefr-vocab. Models are trained on progressively larger subsets of CAES, ranging from 80 samples to the full training size (over 20,000 instances). Figure 4 shows the resulting performance.

The two feature groups exhibit different

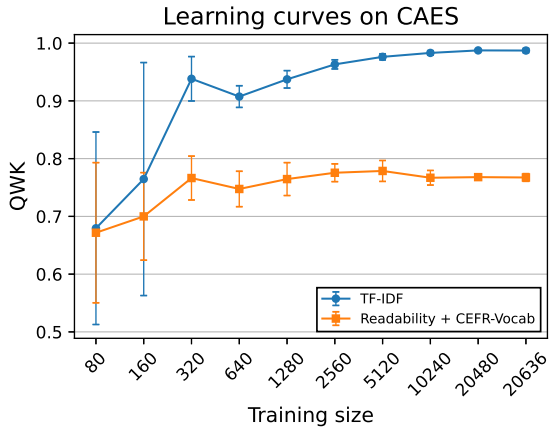

Figure 4: Learning curves on CAES for TF-IDF and readability+cefr-vocab features. Performance is reported as mean QWK across 5-fold cross-validation, with error bars as standard deviation.

scaling behaviour. The TF-IDF configuration improves rapidly with increasing training size, rising from 0.67 at 80 samples to 0.93 at 320 samples, and reaching 0.98 at the full training size. In contrast, readability+cefr-vocab shows only limited gains, increasing from 0.67 at 80 samples to approximately 0.77 at 320 samples, and then remaining almost constant as more training data is added. Therefore, there is a similar performance when less than 100 training samples, but TF-IDF outperforms surface-level features when there are more than 300 samples available for training.

Overall, these results reinforce the findings from the previous experiments. Surface-level (readability+cefr-vocab) provide relatively stable performance in low-resource settings but do not scale with additional data. In contrast, distributional features (TF-IDF) are highly data-dependent, but benefit substantially from increased training size, and achieve higher performance.

5 *Conclusion*

This study examines the effectiveness of different feature representations for CEFR-based text complexity prediction in Spanish as L2. We compared traditional readability metrics, CEFR-aligned lexical overlap features, and distributional representations based on unigram TF–IDF, both individually and in combination. Through a series of experiments, we evaluate their contribution under varying conditions, including cross-dataset generalization, and training size variation.

Regarding RQ1, the results show that distributional features consistently provide the strongest predictive performance when sufficient training data is available, as observed in the CAES corpus. Combining feature groups yields only limited improvements in such settings, indicating that most of the predictive signal is already captured by n-gram representations. In contrast, in smaller datasets, combinations of readability and lexical features become more competitive, suggesting partial complementarity when distributional information is less reliable.

With respect to RQ2, feature representations show limited generalization across datasets. Models trained on learner-produced texts fail to transfer effectively to pedagogical corpora, except for when using TF-IDF features and testing on the smallest corpus (Kwiziq). This suggests that training on larger datasets can provide some cross-dataset benefits, even when the datasets differ. However, this effect is inconsistent and does not generalize across all settings. We leave the exploration of domain adaptation and style normalization strategies as future work, as they may improve transferability and allow the advantages of larger datasets to extend more reliably to smaller ones.

The downsampling experiments demonstrate that differences across feature groups are not only explained by training size, but also reflect corpus-specific properties of the data. If data is limited, surface-level and lexical features provide more stable performance across settings.

These findings have practical implications for Spanish L2 CEFR prediction. In high-resource scenarios, distributional representations are the most effective choice, whereas in low-resource or cross-domain settings, surface-level features offer more stability. Overall, the results highlight the importance of selecting feature representations according to data availability and corpus characteristics, as well as the need for approaches that better capture transferable signals of text complexity.

Future work will explore improved representations of complex lexical structures and CEFR-overlap features for the C1–C2 levels. We also plan to explore transformer-based models with limited training data.

Acknowledgments

This paper presents results from the ESMAS-ES+ project (PID2022-137170OB-I00) funded by MICIU/AEI/10.13039/501100011033 and ERDF/EU.

Iván Arias-Arias acknowledges support from grant FPU21/00188 of the *Formación de Profesorado Universitario* programme, Spanish Ministry of Science, Innovation and Universities.

References

Aissa, W., R. Amaro, D. Antunes, T. Bañeras-Roux, J. Baptista, A. Catala, L. Correia, T. François, M. Garcia, M. Izquierdo-Álvarez, N. Mamede, V. Martins, M. Neves, E. Ribeiro, S. R. Rey, and E. Vanzeveren. 2025. The iRead4Skills Intelligent Complexity Analyzer. In I. Habernal, P. Schulam, and J. Tiedemann, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 73–84, Suzhou, China, November. Association for Computational Linguistics.

Arase, Y., S. Uchida, and T. Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Azpiazu, I. M. and M. S. Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Benedetto, L., G. Gaudeau, A. Caines, and P. Buttery. 2025. Assessing how accurately large language models encode and apply the common european framework of reference for languages. *Computers and Education: Artificial Intelligence*, 8:100353.

Bengoetxea, K. and I. Gonzalez-Dios. 2021. MultiAzterTest: A Multilingual Analyzer on Multiple Levels of Language for Readability Assessment, September.

Bestgen, Y. 2020. Reproducing monolingual, multilingual and cross-lingual CEFR predictions. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5595–5602, Marseille, France, May. European Language Resources Association.

Capel, A. 2015. The English Vocabulary Profile. In J. Harrison and F. Barker, editors, *English Profile in Practice*. Cambridge University Press, Cambridge.

Cohen, J. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213—220, October.

Coniam, D., M. Milanovic, and W. Zhao. 2022. Comparability study between cefr and cse: An exploration using the languagecert test of english. *CEFR Journal - Research and Practice*, 5:25–44, December.

Dale, E. and J. S. Chall. 1948. A Formula for Predicting Readability. *Educational Research Bulletin*, 27(1):11–28.

del Río, I. 2019. Automatic proficiency classification in L2 Portuguese. *Procesamiento del Lenguaje Natural*, 63:–.

Domínguez Vázquez, M. J. 2025. Diseño y metodología de un etiquetador semántico-ontológico multilingüe: ESMAS-ES+. *Revista de Investigación Lingüística*, 28:175–192.

Flesch, R. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

François, T., N. Gala, P. Watrin, and C. Fairon. 2014. FLELex: a graded lexical resource for French foreign learners. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3766–3773, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

He, J. and X. Li. 2024. Zero-shot cross-lingual automated essay scoring. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17819–17832, Torino, Italia, May. ELRA and ICCL.

Hiebert, E. H. 2011. Beyond Single Readability Measures: Using Multiple Sources of Information in Establishing Text Complexity. *Journal of Education*, 191(2):33–42, April.

Imperial, J. M., A. Barayan, R. Stodden, R. Wilkens, R. M. Sanchez, L. Gao, M. Torgbi, D. Knight, G. Forey, R. R. Jablonkai, E. Kochmar, R. Reynolds, E. Ribeiro, H. Saggion, E. Volodina, S. Vajjala, T. François, F. Alva-Manchego, and H. T. Madabushi. 2025. Universalcefr: Enabling open multilingual research on language proficiency assessment.

Lew, R. and S. Wolfer. 2024. CEFR vocabulary level as a predictor of user interest in English Wiktionary entries. *Humanities and Social Sciences Communications*, 11(1):340, February.

Lhoest, Q., A. V. del Moral, Y. Jernite, A. Thakur, P. von Platen, S. Patil, J. Chaumond, M. Drame, J. Plu, L. Tunstall, J. Davison, M. Šaško, G. Chhablani, B. Malik, S. Brandeis, T. L. Scao, V. Sanh, C. Xu, N. Patry, A. McMillan-Major, P. Schmid, S. Gugger, C. Delangue, T. Matussière, L. Debut, S. Bekman, P. Cistac, T. Goehringer, V. Mustar, F. Lagunas, A. M. Rush, and T. Wolf. 2021. Datasets: A community library for natural language processing.

Lively, B. A. and S. L. Pressey. 1923. A method for measuring the "vocabulary burden" of textbooks. *Educational Administration and Supervision*, 9:389–398.

Martinc, M., S. Pollak, and M. Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179, March.

McLaughlin, G. H. 1969. SMOG grading: A new readability formula. *Journal of Reading*, 12(8):639–646.

Miaschi, A., S. Davidson, D. Brunato, F. Dell'Orletta, K. Sagae, C. H. Sanchez-Gutierrez, and G. Venturi. 2020. Tracking the evolution of written language competence in L2 Spanish learners. In J. Burstein, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, and T. Zesch, editors, *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–101, Seattle, WA, USA → Online, July. Association for Computational Linguistics.

Milton, J. 2010. The development of vocabulary breadth across the CEFR levels.

Milton, J. and T. Alexiou. 2009. Vocabulary size and the Common European Framework of Reference for Languages. In B. Richards, H. Daller, D. Malvern, P. Meara, J. Milton, and J. Treffers-Daller, editors, *Vocabulary Studies in First and Second Language Acquisition*. Palgrave Macmillan, Basingstoke, pages 194–211.

Monteiro, R., R. Amaro, S. Correia, A. Pintard, R. Gauchola, M. Moutinho, and X. Blanco Escoda. 2023. iRead4Skills – complexity levels (V1.0).

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12(null):2825–2830, November.

Qi, P., Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In A. Celikyilmaz and T.-H. Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, July. Association for Computational Linguistics.

Ramos, J. 2003. Using tf-idf to determine word relevance in document queries.

In *Proceedings of the First Instructional Conference on Machine Learning*, volume 242, pages 29–48. Citeseer.

Reis, M., S. Barbosa, M. Moutinho, R. Monteiro, S. Correia, and R. Amaro. 2024. Intelligent support for low literacy adults: The European Portuguese iRead4Skills corpus. *International Journal of Emerging Technologies in Learning (iJET)*, 19(08):61–81.

Rey, S. R. and M. Garcia. 2025. Clasificación automática de textos por niveis de lecturabilidade: recursos e modelos para o galego. *Linguamática*, 17(2):33–56, November.

Ribeiro, E., N. Mamede, and J. Baptista. 2024. Automatic Text Readability Assessment in European Portuguese. In P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, and R. Amaro, editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 97–107, Santiago de Compostela, Galicia/Spain, March. Association for Computational Lingustics.

Rodríguez Rey, S. and M. Garcia. 2025. Automatic text readability classification: Resources and models for galician. *Linguamática*, 17(2).

Rodríguez Rey, S. 2024. Automatic text classification using readability levels in galician and spanish. In A. Bonet Jover, D. Vilares Calvo, E. LLoret Pastor, M. Molina-González, and S. Jiménez-Zafra, editors, *Proceedings of the Doctoral Symposium on Natural Language Processing (NLP-DS 2024)*, pages 193–201.

Rojo, G. and I. M. P. Martínez, 2016. *Learner Spanish on computer: The CAES 'Corpus de Aprendices de Español' project*, pages 55–87. John Benjamins Publishing Company.

Schmalz, V. J. and A. Brutti. 2021. Automatic assessment of English CEFR levels using BERT embeddings. In E. Fersini, M. Passarotti, and V. Patti, editors, *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)*, pages 295–301, Milan, Italy, June. CEUR Workshop Proceedings.

Scholz, K. and M. Wenzel. 2025. Evaluating Readability Metrics for German Medical Text Simplification. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6049–6062, Abu Dhabi, UAE, January. Association for Computational Linguistics.

Tanprasert, T. and D. Kauchak. 2021. Flesch-Kincaid is Not a Text Simplification Evaluation Metric. In A. Bosselut, E. Durmus, V. P. Gangal, S. Gehrmann, Y. Jernite, L. Perez-Beltrachini, S. Shaikh, and W. Xu, editors, *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 1–14, Online, August. Association for Computational Linguistics.

Thorndike, E. L. 1921. *The Teacher's Word Book*. Teachers College, Columbia University, New York.

Vajjala, S. 2022. Trends, limitations and open challenges in automatic readability assessment research. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France, June. European Language Resources Association.

Vajjala, S. and T. Rama. 2018. Experiments with universal CEFR classification. In J. Tetreault, J. Burstein, E. Kochmar, C. Leacock, and H. Yannakoudakis, editors, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana, June. Association for Computational Linguistics.

Vásquez-Rodríguez, L., P.-M. Cuenca-Jiménez, S. Morales-Esquivel, and F. Alva-Manchego. 2022. A Benchmark for Neural Readability Assessment of Texts in Spanish. In S. Štajner, H. Saggion, D. Ferrés, M. Shardlow, K. C. Sheang, K. North, M. Zampieri, and W. Xu, editors, *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages

188–198, Abu Dhabi, United Arab Emirates (Virtual), December. Association for Computational Linguistics.

Vogel, M. and C. Washburne. 1928. An objective method of determining grade placement of children's reading material. *The Elementary School Journal*, 28(5):373–381.

Xia, M., E. Kochmar, and T. Briscoe. 2016. Text Readability Assessment for Second Language Learners. In J. Tetreault, J. Burstein, C. Leacock, and H. Yannakoudakis, editors, *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA, June. Association for Computational Linguistics.

Yancey, K. P., G. Laflair, A. Verardi, and J. Burstein. 2023. Rating short L2 essays on the CEFR scale with GPT-4. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, and T. Zesch, editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584, Toronto, Canada, July. Association for Computational Linguistics.

Zalmout, N., H. Saddiki, and N. Habash. 2016. Analysis of Foreign Language Teaching Methods: An Automatic Readability Approach. In H.-H. Chen, Y.-H. Tseng, V. Ng, and X. Lu, editors, *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 122–130, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Zhou, S., H. Jeong, and P. A. Green. 2017. How Consistent Are the Best-Known Readability Equations in Estimating the Readability of Design Standards? *IEEE Transactions on Professional Communication*, 60(1):97–111, March.

A Appendix: Full Results

Dataset	Features	QWK	Accuracy	F1 (weighted)
caes	TF-IDF	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)
caes	CEFR-vocab+TF-IDF	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)
caes	Readability+TF-IDF	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)
caes	All	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)
caes	Readability+CEFR-vocab	0.77 (0.01)	0.62 (0.01)	0.62 (0.01)
caes	Readability	0.70 (0.01)	0.54 (0.00)	0.53 (0.00)
caes	CEFR-vocab	0.63 (0.01)	0.47 (0.00)	0.47 (0.01)
hablacultura	CEFR-vocab+TF-IDF	0.63 (0.07)	0.76 (0.03)	0.75 (0.03)
hablacultura	TF-IDF	0.62 (0.06)	0.76 (0.03)	0.75 (0.04)
hablacultura	Readability+TF-IDF	0.63 (0.07)	0.76 (0.03)	0.75 (0.03)
hablacultura	All	0.63 (0.07)	0.76 (0.03)	0.75 (0.04)
hablacultura	Readability	0.11 (0.10)	0.28 (0.04)	0.30 (0.04)
hablacultura	Readability+CEFR-vocab	0.15 (0.06)	0.27 (0.01)	0.29 (0.01)
hablacultura	CEFR-vocab	0.08 (0.07)	0.19 (0.03)	0.20 (0.02)
kwiziq	Readability+CEFR-vocab	0.51 (0.16)	0.37 (0.03)	0.35 (0.03)
kwiziq	Readability+TF-IDF	0.26 (0.09)	0.34 (0.05)	0.33 (0.06)
kwiziq	CEFR-vocab+TF-IDF	0.25 (0.08)	0.34 (0.05)	0.33 (0.06)
kwiziq	All	0.26 (0.09)	0.34 (0.06)	0.32 (0.06)
kwiziq	TF-IDF	0.25 (0.08)	0.34 (0.05)	0.32 (0.05)
kwiziq	CEFR-vocab	0.23 (0.18)	0.28 (0.03)	0.27 (0.04)
kwiziq	Readability	0.35 (0.09)	0.26 (0.05)	0.24 (0.05)

Table 4: Performance across feature configurations per dataset reported as mean (std) per metric across 5-folds.

Test	Features	QWK	Accuracy	F1 (weighted)
hablacultura	TF-IDF	0.16	0.33	0.35
hablacultura	CEFR-vocab+TF-IDF	0.16	0.32	0.34
hablacultura	Readability+TF-IDF	0.11	0.25	0.28
hablacultura	All	0.11	0.23	0.27
hablacultura	Readability+CEFR-vocab	0.14	0.26	0.24
hablacultura	Readability	0.07	0.22	0.22
hablacultura	CEFR-vocab	0.01	0.08	0.10
kwiziq	CEFR-vocab+TF-IDF	0.33	0.33	0.29
kwiziq	TF-IDF	0.30	0.30	0.27
kwiziq	Readability+TF-IDF	0.31	0.28	0.25
kwiziq	All	0.28	0.28	0.25
kwiziq	Readability	0.34	0.26	0.25
kwiziq	Readability+CEFR-vocab	0.35	0.29	0.24
kwiziq	CEFR-vocab	0.13	0.19	0.19

Table 5: Cross-dataset model performance: trained on CAES and tested on HablaCultura or Kwiziq.

Train size	Features	QWK	Accuracy	F1 (weighted)
650	TF-IDF	0.92 (0.02)	0.91 (0.02)	0.91 (0.02)
650	Readability+TF-IDF	0.91 (0.02)	0.91 (0.03)	0.91 (0.03)
650	CEFR-vocab+TF-IDF	0.92 (0.03)	0.91 (0.03)	0.91 (0.03)
650	All	0.91 (0.02)	0.90 (0.03)	0.90 (0.03)
650	Readability+CEFR-vocab	0.76 (0.06)	0.59 (0.03)	0.58 (0.03)
650	Readability	0.69 (0.05)	0.52 (0.03)	0.51 (0.03)
650	CEFR-vocab	0.67 (0.04)	0.45 (0.02)	0.44 (0.02)
206	Readability+TF-IDF	0.92 (0.05)	0.90 (0.04)	0.90 (0.04)
206	TF-IDF	0.90 (0.07)	0.89 (0.02)	0.89 (0.02)
206	CEFR-vocab+TF-IDF	0.88 (0.09)	0.88 (0.04)	0.88 (0.04)
206	All	0.91 (0.06)	0.88 (0.04)	0.88 (0.04)
206	Readability+CEFR-vocab	0.80 (0.04)	0.54 (0.06)	0.53 (0.04)
206	Readability	0.70 (0.06)	0.50 (0.04)	0.48 (0.03)
206	CEFR-vocab	0.73 (0.02)	0.46 (0.05)	0.45 (0.06)

Table 6: Performance across feature configurations when CAES downsampled reported as mean (std) per metric across 5-folds.

Train size	Features	QWK	Accuracy	F1 (weighted)
80	TF-IDF	0.68 (0.17)	0.72 (0.10)	0.70 (0.11)
80	Readability+CEFR-vocab	0.67 (0.12)	0.45 (0.07)	0.42 (0.07)
160	TF-IDF	0.76 (0.20)	0.81 (0.09)	0.80 (0.10)
160	Readability+CEFR-vocab	0.70 (0.08)	0.47 (0.09)	0.46 (0.10)
320	TF-IDF	0.94 (0.04)	0.88 (0.05)	0.87 (0.05)
320	Readability+CEFR-vocab	0.77 (0.04)	0.57 (0.03)	0.56 (0.03)
640	TF-IDF	0.91 (0.02)	0.91 (0.02)	0.91 (0.02)
640	Readability+CEFR-vocab	0.75 (0.03)	0.59 (0.03)	0.58 (0.03)
1280	TF-IDF	0.94 (0.02)	0.92 (0.01)	0.92 (0.01)
1280	Readability+CEFR-vocab	0.76 (0.03)	0.61 (0.02)	0.61 (0.02)
2560	TF-IDF	0.96 (0.01)	0.95 (0.01)	0.95 (0.01)
2560	Readability+CEFR-vocab	0.78 (0.02)	0.63 (0.02)	0.62 (0.02)
5120	TF-IDF	0.98 (0.01)	0.97 (0.01)	0.97 (0.01)
5120	Readability+CEFR-vocab	0.78 (0.02)	0.63 (0.02)	0.63 (0.02)
10240	TF-IDF	0.98 (0.00)	0.98 (0.00)	0.98 (0.00)
10240	Readability+CEFR-vocab	0.77 (0.01)	0.62 (0.01)	0.62 (0.01)
20480	TF-IDF	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)
20480	Readability+CEFR-vocab	0.77 (0.00)	0.62 (0.01)	0.62 (0.00)
20636	TF-IDF	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)
20636	Readability+CEFR-vocab	0.77 (0.01)	0.62 (0.01)	0.62 (0.01)

Table 7: Learning curve on CAES for TF-IDF and Readability+CEFR-vocab features reported as mean (std) per metric across 5-folds.

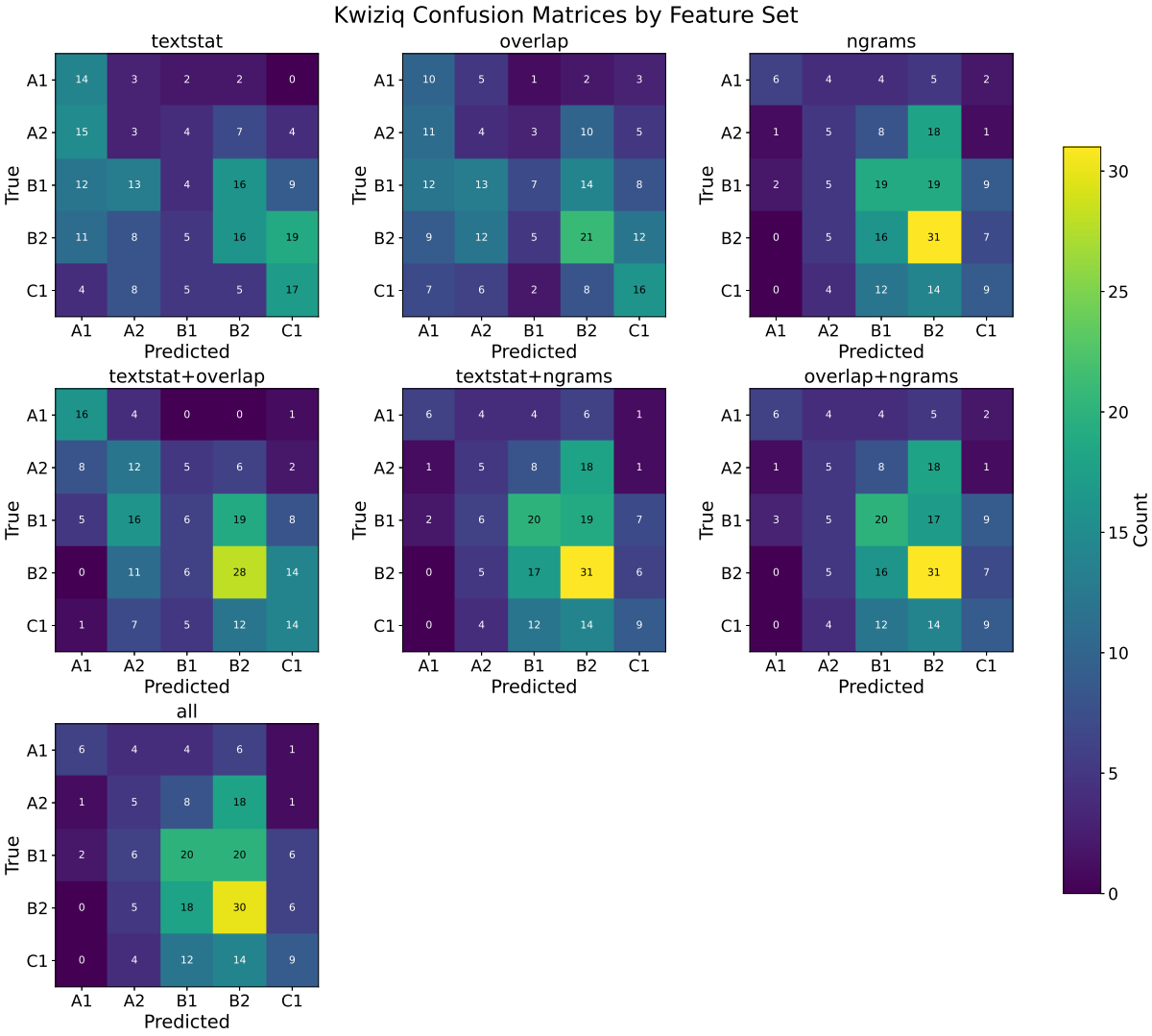

Figure 5: Confusion matrices for predictions on the Kwiziq dataset across 5-fold cross-validation per feature set.