# ETEL: A Lightweight System for Readability Assessment in Basque

## *ETEL: Un Sistema Ligero para la Evaluación de la Legibilidad en Euskera*

**Mikel Iruskieta,**[1]
**Zenón J. Hernández-Figueroa,**[2]**Francisco J. Carreras-Riudavets**[2]
[1]Universidad del Pais Vasco/Euskal Herriko Unibertsitatea
[2]Universidad de Las Palmas de Gran Canaria
[1]mikel.iruskieta@ehu.eus, [2]{zenon.hernandez, francisco.carreras}@ulpgc.es

**Abstract:** Automatic readability assessment is challenging for low-resource languages such as Basque. This paper introduces ETEL, a modular and efficient system for analyzing and classifying Basque texts according to CEFR levels. The system includes a web portal, API, and user management for collaborative corpus creation. ETEL's core uses six computationally inexpensive and interpretable linguistic features. We evaluated these on the 708-text HABE corpus, using statistical analysis to validate their discriminative power and machine learning experiments to benchmark performance. A Random Forest model using ETEL's features achieves 0.83 accuracy in a 3-level (Basic/Intermediate/Advanced) classification and a competitive 0.93 accuracy in a binary (Basic/Advanced) task. This work provides preliminary empirical evidence supporting the effectiveness of ETEL as a lightweight tool for Basque readability assessment, offering a valuable resource for educators and a methodological blueprint for other low-resource languages. The system was developed by two nodes of the CLARIAH-ES infrastructure.
**Keywords:** Basque language processing, Text classification, Text readability assessment, Text complexity analysis.

**Resumen:** La evaluación automática de la legibilidad es un desafío para lenguas con escasos recursos, como el euskera. Este artículo presenta ETEL, un sistema modular y eficiente para analizar y clasificar textos en euskera según los niveles del MCER. El sistema incluye un portal web, una API y gestión de usuarios para la creación colaborativa de corpus. El núcleo de ETEL utiliza seis características lingüísticas computacionalmente económicas e interpretables. Las evaluamos en el corpus HABE de 708 textos, utilizando análisis estadístico para validar su poder discriminativo y experimentos de aprendizaje automático para medir su rendimiento. Un modelo Random Forest con las características de ETEL alcanza una precisión de 0.83 en una clasificación de 3 niveles (Básico/Intermedio/Avanzado) y una precisión competitiva de 0.93 en una clasificación binaria (básico/avanzado). Este trabajo aporta evidencia empírica preliminar que respalda la eficacia de ETEL como herramienta ligera para la evaluación de la legibilidad en euskera, ofreciendo un recurso valioso para educadores y un modelo metodológico para otras lenguas con pocos recursos. El sistema fue desarrollado por dos nodos de la infraestructura CLARIAH-ES.
**Palabras clave:** Procesamiento del lenguaje vasco, Clasificación de textos, Evaluación de la legibilidad de textos, Análisis de la complejidad de textos.

## 1 Introduction

The automatic assessment of text readability is a long-standing goal in Natural Language Processing (NLP), with profound implications for education, content creation, and second language acquisition (Vajjala, 2022). By automatically assigning a difficulty level to a text, systems can help teachers select appropriate reading materials, guide students to suitable content, and support writers in tailoring their texts to a specific audience. For well-resourced languages like English, numerous sophisticated tools and datasets exist (McNamara et al., 2014)(Collins-Thompson, 2014). However, the challenge is significantly greater for low-resource languages, which lack the large, annotated corpora and foundational NLP tools necessary to train complex models (Vajjala and Rama, 2018).

Basque, a language isolate spoken by approximately 800,000 people, is a prime example of such a low-resource language. While progress has been made in developing basic NLP tools, resources for higher-level tasks like readability assessment remain scarce. This scarcity creates a critical need for computationally efficient systems that can perform reliably without depending on massive datasets or complex deep learning architectures.

This paper introduces ETEL[1] (Euskarazko Testuen Ezaugarri Linguistikoa), a new, easy-to-use system designed to analyze texts in Basque and obtain linguistic complexity measures that allow classifying them within a scale of complexity levels. The main contributions of this work are threefold:

1. **System Presentation:** We present the complete architecture of the ETEL system, a modular and scalable platform that includes a web portal, a bulk file uploader, an analysis API, and a user management system with different roles to facilitate collaborative corpus building. The system is designed to be a reusable and adaptable infrastructure for readability research.
2. **Feature Validation:** We propose and validate a minimal set of six computationally inexpensive linguistic features for Basque. We conduct a rigorous statistical analysis on a 708-text corpus, demonstrating that these features have a good discriminative power across most CEFR levels.
3. **Comprehensive Evaluation:** We perform a series of robust machine learning experiments to benchmark the classification performance of ETEL. We show that our lightweight approach achieves robust performance in practical scenarios (3-level classification) and is competitive with a much more complex state-of-the-art system in a binary classification task.

By presenting both the system and its comprehensive evaluation, we provide preliminary empirical evidence that ETEL is an effective and practical approach to readability assessment in Basque, providing a valuable tool for educators and researchers, and a methodological framework for similar efforts in other low-resource contexts.

## 2 Related work

Automatic Readability Assessment (ARA) has evolved significantly from early, formula-based approaches to modern, data-driven machine learning models. This section situates our work within the broader context of ARA, with a particular focus on the challenges and recent advancements in low-resource languages.

### 2.1 From Readability Formulas to Feature-Based Machine Learning

Early approaches to ARA were dominated by readability formulas such as the Flesch-Kincaid Grade Level (Kincaid et al., 1975) and Gunning Fog Index (Gunning, 1952). These formulas rely on simple surface-level features, typically word length (in characters or syllables) and sentence length (in words). While easy to compute, they have been widely criticized for their linguistic naivety, as they ignore crucial factors like syntactic complexity, lexical semantics, and discourse structure (Benjamin, 2012).

To overcome these limitations, the field shifted towards feature-based machine learning models. This paradigm, which represents the mainstream approach for the last two decades, involves engineering a wide range of linguistic features to train supervised classifiers. Systems like Coh-Metrix for English

[1] ETEL is available at https://etel.iatext.ulpgc.es/

(McNamara et al., 2014) exemplify this approach, often using hundreds of features to achieve high accuracy.

### 2.2 Readability Assessment in Low-Resource Languages

The development of ARA tools for low-resource languages faces significant hurdles, primarily the scarcity of annotated corpora and the lack of robust, underlying NLP pipelines (parsers, taggers) needed for feature extraction (Vajjala and Rama, 2018). Research in this area often follows one of two paths: (1) adapting existing formulas, or (2) building feature-based models with whatever linguistic tools are available.

Recent work on other low-resource European languages highlights these challenges. For instance, a study on Estonian learner texts (A2-C1) used a combination of lexical, morphological, and error-based features to achieve a high accuracy of 0.9, demonstrating the power of careful feature engineering even without deep learning (Allkivi, 2026). Similarly, a project for Galician and Spanish aimed to create new corpora and evaluate computational strategies for readability, underscoring the foundational need for data (Rodriguez Rey, 2024). ETEL follows this second path: it builds a feature-based classification model using a minimal set of computationally inexpensive linguistic indicators extractable with the NLP tools currently available for Basque, without relying on large annotated corpora or deep learning architectures.

### 2.3 The Rise of Neural Models and Large Language Models (LLMs)

The last few years have seen a paradigm shift in NLP with the advent of large-scale pre-trained language models like BERT (Devlin et al., 2019) and, more recently, Large Language Models (LLMs) such as GPT-4 (OpenAI, 2023). These models have also been applied to ARA.

Initial approaches used pre-trained embeddings from models like BERT as features for a classifier, often outperforming traditional feature-based systems, especially when training data is abundant (Lee and Vajjala, 2022). More recent work has explored the use of LLMs in zero-shot or few-shot settings, where the model is prompted to directly predict the readability level of a text. The ReadMe++ benchmark, a multilingual dataset covering five languages, showed that while LLMs are powerful, they do not consistently outperform fine-tuned, smaller models, especially in domain-specific contexts (Naous et al., 2024).

A key challenge for LLMs is their applicability to low-resource languages like Basque. While multilingual models have some capability, their performance is often degraded. A recent study on Automatic Essay Scoring for Basque demonstrated that a locally fine-tuned model (Latxa) could outperform massive proprietary models like GPT-5 in scoring consistency and feedback quality, highlighting the continued importance of language-specific adaptation (Azurmendi, Arregi, and Lopez de Lacalle, 2025).

For Basque, the landscape is defined by two main systems prior to our work: ErreXail (Gonzalez-Dios, Aranzabe, and Díaz de Ilarraza, 2014) and MultiAzterTest (Bengoetxea and Gonzalez-Dios, 2021). ErreXail was a pioneering system that used 94 linguistic indices for binary (simple/complex) classification. MultiAzterTest expanded on this, creating a multilingual tool that also performs binary classification for Basque, reporting an accuracy of around 95.5%. Both systems rely on many complexes, engineered features. Our study consciously adopts a "pre-LLM" feature-based approach for two strategic reasons. First, it allows for a direct and fair evaluation of the ETEL system's design principles. Second, it advocates a methodology that is interpretable, computationally frugal, and not reliant on massive, expensive-to-train models. We demonstrate that this "classic" approach remains highly effective and relevant, particularly in low-resource contexts where efficiency and explainability are paramount.

## *3 The ETEL System*

ETEL is a modular system designed for scalability, reusability, and adaptation to other languages. This section describes its architecture and core functionalities.

### 3.1 System Architecture

The system has been developed using Docker and integrates both proprietary and third-party services. Figure 1 shows the interaction between the web front-end, back-end ser-

vices, and third-party components such as MySQL and MinIO. The proprietary services developed are:

- **ETEL-Web:** The main web portal where users interact with the system's functionalities.
- **ETEL-FileUploader:** is a console application that allows for the mass upload of files for corpus building.
- **ETEL-Analyzer:** A web API that exposes services to analyze plain text or a file, returning its linguistic metrics as a JSON response.
- **ETEL-Shared:** A class library containing shared utilities to ensure code consistency across the different services.

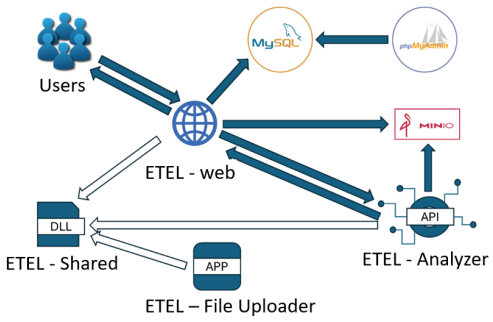


Figure 1: The modular architecture of the ETEL system.

These services are supported by standard third-party components:

- **MySQL:** The database engine for storing user information, file metadata, and calculated metrics.
- **PhpMyAdmin:** A graphical client for database management.
- **MinIO:** An S3-compatible object storage service for physically storing the text files.

### 3.2 User Roles and Functionality

ETEL is designed as a collaborative platform and supports four distinct user roles:

- **Anonymous User:** Can analyze short texts (up to 2500 characters) and view public corpora.
- **Researcher/Contributor:** Can upload texts to a corpus, classify them by complexity level, and manage their own contributions.
- **Administrator:** Can create new corpora, define complexity scales, manage access restrictions, and invite other users to be contributors.
- **Superadministrator:** Has full control over user management, including creating, deleting, and assigning roles.

This role-based system allows for both public use and controlled, collaborative research, making ETEL a flexible tool for the language technology community.

## *4 Experimental Evaluation*

### 4.1 The HABE Corpus

To validate the effectiveness of ETEL's lightweight approach, we conducted a comprehensive experimental evaluation using the HABE corpus, a collection of 708 Basque texts manually classified by language experts into the six CEFR levels. This corpus has been provided by HABE (Helduen Alfabetatze eta Berreuskalduntzerako), and the texts have been created for Ikasbil, which offers reading texts and exercises for all levels according to the CEFR. The distribution of texts across these levels is shown in Figure 2. The pie chart shows the severe class imbalance, with the majority of texts concentrated in the B1 (45.2%) and B2 (37.1%) levels. It is important to note that all texts in this corpus are pedagogical materials explicitly curated for language learners. This educational nature conditions the distribution of linguistic features—for instance, advanced texts are systematically longer—which may limit the model's direct generalizability to spontaneous, non-pedagogical domains.

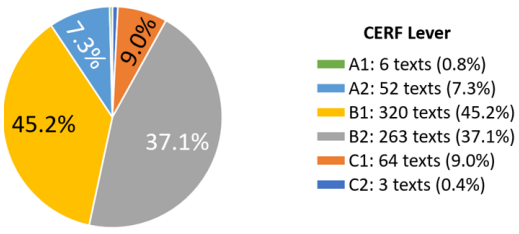


Figure 2: Distribution of the 708 texts in the HABE corpus across the six CEFR levels.

## 4.2 Linguistic Features

Before selecting this final set, we considered incorporating deeper morphological and syntactic features (e.g., case suffixes or noun phrase complexity indicators), given the agglutinative nature of Basque. However, they were ultimately excluded from this initial version because they either required NLP pipelines that were too computationally expensive for our lightweight design goals, or the available tools did not offer sufficient reliability for robust feature extraction across all proficiency levels. Consequently, the selected features are:

- **NWT (Number of Word Tokens):** The total number of words in the text. A proxy for text length.
- **NSy (Number of Syllables):** The total number of syllables. Related to word and sentence length.
- **NPFVLT (Number of POS Feature: Verb Lemma Types):** The number of unique verb lemmas. This metric aims to capture the diversity of verbal actions in the text.
- **TTRU10LW (Type/Token Ratio – Uber10 Lexical Words):** A measure of lexical diversity focusing on the 10% least frequent lexical words, calculated as
$$\frac{\log_{10}(nlw)^2}{\log_{10}(nlw) - \log_{10}(ndlw)} \quad (1)$$
nlw = number of lexical words
ndlw = number of different lexical words
- **CTTR (Corrected Type/Token Ratio):** A variant of TTR that aims to be more robust to variations in text length, calculated as Types / $\sqrt{2 * \text{Tokens}}$.
- **CTTRLEM (Corrected Type/Token Ratio – Lemmas):** The same as CTTR but calculated on lemmas instead of word forms.

## 4.3 Experimental design

To provide a robust evaluation, we designed a multi-stage experimental process.

### 4.3.1 Statistical Analysis of Features

First, to determine if the six ETEL features have a statistically significant relationship with the CEFR levels, we performed a Kruskal-Wallis H-test for each feature. This non-parametric test is suitable for comparing more than two independent groups and does not assume a normal distribution of the data. For features showing a significant global difference, we then conducted a Dunn's post-hoc test with Bonferroni correction to identify which specific pairs of CEFR levels (e.g., A1 vs. A2, B1 vs. C1) showed significant differences.

### 4.3.2 Classification Experiments

We framed the readability assessment as a supervised machine learning task. We conducted experiments for three distinct classification scenarios:

- **6-Level Classification:** A fine-grained task to predict the specific CEFR level (A1-C2).
- **3-Level Classification:** A coarse-grained task where levels are grouped into Basic (A1, A2), Intermediate (B1, B2), and Advanced (C1, C2). This is a more practical scenario given the corpus imbalance.
- **2-Level Classification:** Basic (A1, A2) and Advanced (B1, B2, C1, C2). This specific grouping—where the threshold B1 level is considered "Advanced"—was chosen to align with the "simple vs. complex" binary definition used by MultiAzterTest, enabling a direct comparison. Furthermore, within the context of the HABE corpus, texts from B1 upwards are designed for learners with autonomous communicative competence, whereas A1-A2 texts target initial learning stages.

For the three scenarios, we evaluated a set of four standard, yet diverse, classification algorithms:

- **Logistic Regression:** A robust linear baseline.
- **Support Vector Machine (SVM):** A powerful non-linear model using a Radial Basis Function (RBF) kernel.
- **Random Forest:** An ensemble of decision trees, known for its high performance and robustness.
- **Gradient Boosting:** Another powerful tree-based ensemble method that builds trees sequentially.

| Feature | A1 | A2 | B1 | B2 | C1 | C2 |
|---|---|---|---|---|---|---|
| TTRU10LW | 62.12 ± 12.09 | 58.20 ± 28.66 | 54.67 ± 27.19 | 58.94 ± 23.05 | 73.83 ± 67.33 | 57.72 ± 4.27 |
| NWT | 251.00 ± 20.21 | 415.65 ± 363.11 | 646.56 ± 322.33 | 717.25 ± 375.40 | 803.83 ± 499.97 | 1962.00 ± 1256.64 |
| NSy | 676.50 ± 73.46 | 1157.52 ± 1009.62 | 1835.42 ± 907.69 | 2062.50 ± 1013.87 | 2338.77 ± 1382.75 | 5438.33 ± 3647.52 |
| NPFVLT | 45.67 ± 4.46 | 81.23 ± 73.54 | 120.95 ± 65.19 | 129.92 ± 77.30 | 142.58 ± 94.47 | 380.33 ± 217.09 |
| CTTR | 7.97 ± 0.49 | 8.30 ± 2.78 | 10.53 ± 1.96 | 11.27 ± 1.52 | 12.04 ± 2.78 | 17.39 ± 4.74 |
| CTTRLEM | 6.28 ± 0.62 | 6.02 ± 1.61 | 7.44 ± 1.18 | 8.00 ± 1.04 | 8.78 ± 1.80 | 11.69 ± 3.53 |

Table 1: Descriptive statistics of ETEL linguistic features by CEFR level (Mean ± Standard Deviation; corpus HABE, N = 708 texts in Basque).

All experiments were conducted using a stratified 5-fold cross-validation protocol. We do not use stratified 10-fold cross-validation, although it is a widely accepted evaluation protocol for supervised classification on datasets of moderate size, because, for this particular corpus, the minority class is too small: the gain in estimation accuracy would be negated by the instability introduced by having too few examples from the most underrepresented CEFR levels, particularly A1 and C2 in the 6-level task, within individual evaluation folds. A more robust balance is achieved with 5 folds. Stratification ensures that the class distribution in each fold mirrors the distribution in the overall dataset, which is crucial for imbalanced corpora. Performance was measured using Accuracy, Macro-F1 score, Weighted-F1 score, and Cohen's Kappa ($\kappa$).

#### 4.3.3 Handling Class Imbalance

To address the significant class imbalance (e.g., A1 has 6 texts, while B1 has 320), we applied class weighting (`class_weight='balanced'`) where supported by the algorithm. We did not try using SMOTE because, with classes as small as A1 and C2, it does not expand the actual variability of the class; it simply fills a very narrow space artificially, leading the model to learn a decision boundary based on data that do not represent the true diversity of the class.

## *5 Results*

### 5.1 Descriptive and Statistical Analysis of Features

We first analyzed the distribution of each of the six features across the CEFR levels. As shown in Table 1, all features exhibit a clear trend of increasing values with higher proficiency levels, with the exception of TTRU10LW, which shows a less consistent pattern. The high standard deviations in levels like A and C reflect the internal heterogeneity of texts within those levels.

| Metric | H-statistic | p-value | Significant |
|---|---|---|---|
| NWT | 54.26 | 1.85e-10 | Yes |
| NSy | 59.28 | 1.72e-11 | Yes |
| NPFVLT | 41.84 | 6.35e-08 | Yes |
| CTTR | 93.11 | 1.49e-18 | Yes |
| CTTRLEM | 121.86 | 1.27e-24 | Yes |
| TTRU10LW | 19.75 | 1.39e-03 | Yes |

Table 2: Results of the Kruskal-Wallis H-test for each linguistic feature across the six CEFR levels.

To formally test the discriminative power of each feature, we performed a Kruskal-Wallis H-test which confirmed these visual trends. As shown in Table 2, all six features were found to be statistically significant, indicating that they can reliably distinguish between at least some of the CEFR levels. CTTRLEM emerged as the most discriminative feature (H=121.86), followed by CTTR (H=93.1). The TTRU10LW feature has the lowest H statistic (19.75), which suggests a lower discriminatory power between levels.

Dunn's post-hoc tests (Figure 3) provide a more granular view, revealing which specific level pairs are distinct. CTTRLEM is the feature that discriminates between the most pairs of levels, showing significant differences between A2 and all higher levels (B1, B2, C1, C2), between B1 and B2, between B1 and C1, and between B2 and C1. This makes it the most informative metric for differentiating CEFR levels. CTTR exhibits a similar pattern, with significant differences between A2 and all levels B1-C2, and between B1 and B2, and between B1 and C1. NWT and NSy discriminate well between lower levels (A1, A2) and intermediate-advanced levels (B1-C2), but not between adjacent levels within B1-C2. TTRU10LW is the feature with the lowest discriminatory capacity in pairwise comparisons, showing a significant difference only between B1 and C1.

These results characterise the univariate discriminative power of each feature. In the

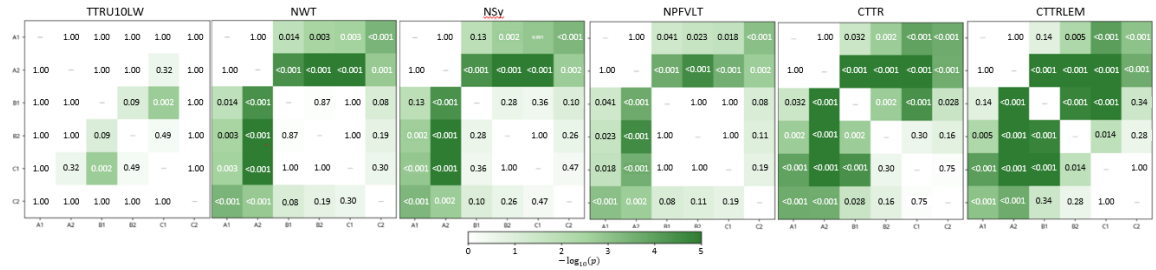


Figure 3: Heatmaps of p-values from Dunn's post-hoc test with Bonferroni correction for each feature.

next section we complement this view with a multivariate analysis based on Random Forest feature importance, which may yield a different ranking due to interactions and redundancy between predictors.

## 5.2 Classification Performance

For the 6-level classification task (Table 3), Gradient Boosting achieves the highest raw accuracy (0.513) and the highest $\kappa$ (0.207), but its Macro-F1 (0.318) is equal to the SVM and lower than Random Forest. This is because Gradient Boosting, unlike Random Forest, does not support `class_weight='balanced'`; therefore, it tends to focus on the majority classes (B1, B2), inflating accuracy and $\kappa$ while neglecting the minority classes, as reflected in the lower Macro-F1. Random Forest achieves the highest Macro-F1 among all classifiers (0.363), meaning it distributes its predictive power more evenly across all six levels, including the minority classes (A1 with 6 texts, C2 with 3 texts). The higher standard deviation ($\pm$0.039 for accuracy, $\pm$0.078 for $\kappa$) reflects sensitivity to how the rare classes fall in each fold. The non-linear kernel allows the SVM to capture more complex decision boundaries than Logistic Regression. Accuracy improves by 10 percentage points and $\kappa$ nearly doubles, indicating a meaningfully better agreement with the true labels. The low standard deviations ($\pm$0.011 for accuracy) show that this model is very stable across folds. Logistic Regression performs worst in the 6-level task. Its accuracy (0.298) is substantially below the majority-class baseline of approximately 0.45, which would be obtained by always predicting B1, the most frequent class in the corpus. This result shows that the model struggles to learn a useful decision boundary under extreme class imbalance, despite the use of class weighting. The low Cohen's Kappa (0.091) is consistent with this interpretation.

| Classifier | Accuracy | Macro-F1 | Weighted-F1 | Kappa ($\kappa$) |
|---|---|---|---|---|
| Logistic Regression | 0.298 ± 0.019 | 0.253 ± 0.035 | 0.323 ± 0.021 | 0.091 ± 0.026 |
| SVM (RBF) | 0.394 ± 0.011 | 0.318 ± 0.072 | 0.410 ± 0.008 | 0.150 ± 0.022 |
| Random Forest | 0.510 ± 0.039 | 0.363 ± 0.068 | 0.495 ± 0.043 | 0.195 ± 0.078 |
| Gradient Boosting | 0.513 ± 0.019 | 0.318 ± 0.024 | 0.501 ± 0.017 | 0.207 ± 0.034 |

Table 3: Full classification results for the 6-level task (Mean ± SD from 5-fold stratified CV).

| Classifier | Accuracy | Macro-F1 | Weighted-F1 | Kappa ($\kappa$) |
|---|---|---|---|---|
| Logistic Regression | 0.552 ± 0.041 | 0.449 ± 0.030 | 0.609 ± 0.039 | 0.205 ± 0.048 |
| SVM (RBF) | 0.630 ± 0.046 | 0.492 ± 0.036 | 0.677 ± 0.034 | 0.237 ± 0.039 |
| Random Forest | 0.831 ± 0.012 | 0.500 ± 0.051 | 0.796 ± 0.014 | 0.250 ± 0.067 |
| Gradient Boosting | 0.811 ± 0.008 | 0.497 ± 0.041 | 0.788 ± 0.012 | 0.242 ± 0.054 |

Table 4: Full classification results for the 3-level task (Mean ± SD from 5-fold stratified CV).

| Classifier | Accuracy | Macro-F1 | Weighted-F1 | Kappa ($\kappa$) |
|---|---|---|---|---|
| Logistic Regression | 0.780 ± 0.032 | 0.609 ± 0.029 | 0.825 ± 0.024 | 0.260 ± 0.046 |
| SVM (RBF) | 0.845 ± 0.015 | 0.658 ± 0.026 | 0.869 ± 0.010 | 0.331 ± 0.053 |
| Random Forest | 0.931 ± 0.010 | 0.675 ± 0.086 | 0.916 ± 0.018 | 0.360 ± 0.166 |
| Gradient Boosting | 0.915 ± 0.013 | 0.647 ± 0.066 | 0.904 ± 0.016 | 0.298 ± 0.130 |

Table 5: Full classification results for the 2-level task.

Performance improves substantially as the classification task is simplified, as evidenced by comparing the results across the 6-level, 3-level, and 2-level scenarios (Table 3, 4, and 5). The most dramatic gain occurs in the transition from 6 to 3 levels, where the Random Forest accuracy jumps from 0.510 to 0.831. This reflects that grouping adjacent CEFR levels into broader categories (e.g., B1+B2 into 'Intermediate') eliminates the most frequent confusion pairs.

In the 3-level task, Random Forest achieves an accuracy of 0.831 compared to 0.810 for Gradient Boosting, with both reaching similar Macro-F1 scores. The further simplification to a binary setting (Basic vs. Advanced) pushes Random Forest accuracy to 0.931, although the Macro-F1 gain is more modest (0.500 $\rightarrow$ 0.675) because the Basic

class remains difficult to recover due to its limited size (58 texts). Across all scenarios, the narrowing gap between Macro-F1 and Weighted-F1 confirms that class imbalance becomes less damaging in coarser schemes. Furthermore, the fact that tree-based ensembles consistently outperform linear models (Logistic Regression and SVM) suggests that the underlying relationships between ETEL metrics and proficiency levels are inherently non-linear.

Figure 4 presents the confusion matrices for the Random Forest classifier on the 6-level CEFR classification task, with the left panel showing absolute prediction counts and the right panel showing row-normalised proportions. The most salient structural pattern is that most misclassifications occur between adjacent CEFR levels, particularly between B1 and B2, although some non-adjacent errors are also observed. This suggests that the six ETEL features capture the overall progression of linguistic complexity across the scale, while still showing limited discriminative resolution for some minority and upper-level classes. The majority classes B1 and B2 dominate the absolute counts, yet even they show substantial mutual confusion: 46% of true B2 texts are predicted as B1, reflecting the inherent linguistic proximity of these two neighbouring levels. The minority classes suffer the most: A1 (6 texts) achieves a recall of only 17%, with half of its texts misclassified as B1, while A2 (52 texts) reaches just 25% recall, with 62% of its texts absorbed by the B1 category. C1 fares poorly as well, with only 20% correct recall and the majority of its texts confused with B2 (48%) and B1 (30%), suggesting that the ETEL features — which primarily capture text length and vocabulary diversity — do not sharply discriminate higher intermediate levels from advanced ones. C2, with only 3 texts, achieves 33% recall, though errors remain within the adjacent C1 level, which is a relatively coherent outcome given the extreme data scarcity. The gap between Weighted-F1 (0.50) and Macro-F1 (0.36) observed in Table 3 is directly explained by this figure: the model performs reasonably on the dominant B1/B2 classes while failing on the extremes of the scale. Overall, Figure 4 confirms that the 6-level task is fundamentally constrained by both the severe class imbalance of the HABE corpus and the limited discriminative resolution of six surface-level features, motivating the adoption of coarser 3-level and 2-level classification schemes.

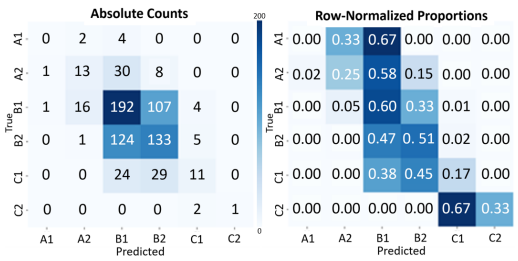


Figure 4: Confusion matrices for the 6-level classification task using Random Forest.

### 5.3 Feature importance

Figure 5 displays the Gini importance scores assigned by the Random Forest model to each of the six ETEL features in the 6-level classification task, ranked from most to least discriminative. The distribution of scores shows that all six features contribute meaningfully to the model, with no single feature dominating it, although the range is not uniform: CTTR leads at 0.197 while TTRU10LW trails at 0.111, a difference of approximately 44%. CTTR (Corrected Type/Token Ratio) emerges as the most informative feature (0.197), followed closely by NWT (Number of Word Tokens, 0.184), NSy (Number of Syllables, 0.176), and NPFVLT (Number of Verb Lemma Types, 0.172). This cluster of four top features reflects two complementary dimensions of text complexity: lexical richness (CTTR captures vocabulary diversity relative to text length) and text volume (NWT, NSy, and NPFVLT all grow as texts become longer and more linguistically elaborate at higher CEFR levels). CTTRLEM (CTTR computed on lemmas, 0.159) ranks fifth, offering a slightly less discriminative variant of the same lexical diversity signal already captured by CTTR. TTRU10LW (Type/Token Ratio – Uber10 Lexical Words, 0.111) is the least important feature, suggesting that focusing on the 10% least frequent lexical words provides less additional discriminative power once the other features are already present in the model. While all six features contribute to the model, their importance scores are not evenly distributed: the top four features (CTTR, NWT, NSy, NPFVLT) cluster between 0.172 and 0.197, while CTTRLEM and especially TTRU10LW contribute

noticeably less. This spread suggests that TTRU10LW adds limited marginal information once the other features are present, and that CTTR and CTTRLEM partially overlap in the lexical diversity signal they encode. Removing any single feature would therefore be expected to degrade performance to some extent, but not necessarily to the same degree for all predictors. When compared with the Kruskal–Wallis results, the Random Forest importance analysis offers a complementary, multivariate perspective. Kruskal–Wallis evaluates each feature in isolation, and therefore ranks CTTRLEM as the most discriminative predictor, slightly ahead of CTTR. In contrast, the Random Forest relies more heavily on CTTR once both metrics are present in the same model, and assigns a lower marginal importance to CTTRLEM, which suggests that much of the lexical diversity signal carried by CTTRLEM is already captured by CTTR. This difference in ranking is thus expected: it reflects interactions and redundancy between features rather than a contradiction between the two analyses.

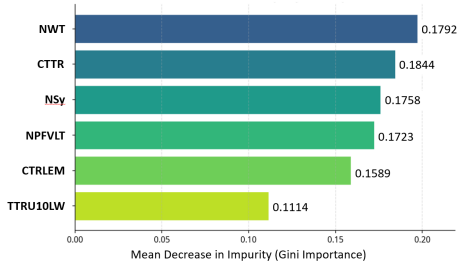


Figure 5: Feature importance scores - Random Forest (6-Level Classification).

### 5.4 Comparison with State-of-the-Art

We compared the effectiveness of ETEL features in the Basic/Advanced classification with the results obtained with MultiAzterTest, the primary classification tool available for Basque. To do this, we installed MultiAzterTest locally using the version available on GitHub[2] and ran it on the 708 files of the HABE corpus. Although the comparison with MultiAzterTest is informative, it should be interpreted cautiously because the two systems were evaluated under different protocols and therefore do not constitute a strictly equivalent benchmark. Specifically, ETEL's results are obtained via 5-fold stratified cross-validation, which estimates out-of-sample generalisation performance. MultiAzterTest, by contrast, was run in direct prediction mode on the full 708-text corpus without a held-out test set, meaning its reported metrics reflect in-sample performance. Consequently, the apparent advantage of ETEL should be interpreted as indicative rather than conclusive.

The MultiAzterTest achieved an overall accuracy of 92.2%, slightly lower than that the 95.5% reported in (Bengoetxea and Gonzalez-Dios, 2021), attributable to the differences between the HABE corpus and the one used in the original study. A closer inspection reveals a strong bias towards the majority class ("Advanced"). It correctly identified nearly all Advanced texts (Recall = 1.00 for Advanced) but failed significantly on the Basic texts (Recall = 0.10 for Basic). This results in a low Macro-F1 score of 0.569 and a Cohen's Kappa of 0.161.

In contrast, the Random Forest model trained on just six ETEL features achieved a higher accuracy (0.931) and a better Macro-F1 score (0.675) and Cohen's Kappa (0.360) during cross-validation, demonstrating better handling of the minority "Basic" class despite the severe dataset imbalance.

| System | Accuracy | Macro-F1 | Weighted-F1 | Kappa ($\kappa$) |
|---|---|---|---|---|
| ETEL – Logistic Regression | 0.780 ± 0.032 | 0.609 ± 0.029 | 0.825 ± 0.024 | 0.260 ± 0.046 |
| ETEL – SVM (RBF) | 0.845 ± 0.015 | 0.658 ± 0.026 | 0.869 ± 0.010 | 0.331 ± 0.053 |
| ETEL – Random Forest | 0.931 ± 0.010 | 0.675 ± 0.086 | 0.916 ± 0.018 | 0.360 ± 0.166 |
| ETEL – Gradient Boosting | 0.915 ± 0.013 | 0.647 ± 0.066 | 0.904 ± 0.016 | 0.298 ± 0.130 |
| MultiAzterTest (94 metrics) | 0.922 | 0.569 | 0.895 | 0.161 |

Table 6: Comparison Table: Binary Classification (Basic vs. Advanced).

The analysis suggests that a parsimonious model using only six linguistic features (ETEL) combined with a robust algorithm like Random Forest (with class weighting) can compete with a highly complex tool (MultiAzterTest with 94 metrics) in classifying the readability of Basque texts, particularly in mitigating the effects of class imbalance to accurately identify basic-level texts (Table 6).

## *6 Conclusions and Future Work*

In this paper, we introduced ETEL, a lightweight, modular, and efficient system designed for automatic readability assessment in Basque. Given the scarcity of resources for Basque NLP, our approach deliberately eschews massive, computationally expensive

[2]https://github.com/kepaxabier/MultiAzterTest

models in favor of a parsimonious, feature-based architecture. The core of ETEL relies on just six computationally inexpensive linguistic features features —such as Corrected Type/Token Ratio (CTTR) and Number of Word Tokens (NWT)— which we have demonstrated to be consistently discriminative across CEFR proficiency levels.

Through a rigorous statistical analysis on the 708-text HABE corpus, we empirically demonstrated the discriminative power of these features, confirming that lexical diversity and text volume are strong indicators of linguistic complexity in Basque. Our machine learning experiments further corroborated the effectiveness of this lightweight approach. A Random Forest classifier utilizing only ETEL's six features achieved a competitive accuracy of 0.931 and a Macro-F1 score of 0.675 in a binary (Basic vs. Advanced) classification task. Notably, this parsimonious model showed better performance than MultiAzterTest in handling class imbalance and identifying minority Basic texts; however, this comparison should be interpreted cautiously because the two systems were evaluated under different protocols and therefore do not constitute a strictly equivalent benchmark.

While the system faced challenges in the fine-grained 6-level classification task, primarily due to the severe class imbalance in the training corpus and the inherent difficulty of distinguishing between adjacent intermediate levels, its performance in 3-level and binary classifications confirms its practical utility. ETEL provides a valuable, accessible tool for educators and content creators working with the Basque language, while also offering a methodological blueprint that prioritizes efficiency and interpretability for other low-resource languages.

Despite the promising results, several avenues remain open for future research and system enhancement:

Firstly, the most pressing limitation is the severe class imbalance and limited size of the HABE corpus, particularly regarding the extreme proficiency levels (A1 and C2). Future efforts must focus on expanding the corpus through collaborative data collection using the ETEL platform's user management and corpus-building features. Beyond opening the platform to collaborative user contributions, we plan to actively address this by establishing agreements with Basque educational institutions to collect verified learner productions, and by exploring controlled data augmentation techniques (e.g., expert-supervised paraphrasing) for the most scarce levels. A more balanced dataset is essential for improving the model's performance in fine-grained, multi-class scenarios.

While the HABE corpus provides a standardized benchmark for educational materials, future research should focus on cross-domain validation. Testing ETEL on spontaneous linguistic productions, such as news articles, social media, or literary texts, will be essential to assess the system's robustness and generalizability beyond the curated pedagogical context. Since spontaneous productions like daily journalism or social media introduce drastically different syntactic structures, the current model's reliance on volume-based features (NWT, NSy) might lead to misclassifications if the correlation between text length and complexity—inherent to the HABE pedagogical design—does not hold in open domains.

Secondly, while the current set of six surface-level features is efficient, incorporating a limited number of deeper linguistic features could improve discrimination between adjacent proficiency levels. Specifically, exploring computationally light syntactic or morphological complexity metrics—tailored to the agglutinative nature of Basque—could provide the necessary signal to better differentiate intermediate and advanced texts without compromising the system's lightweight design.

Finally, we plan to conduct user studies with educators and language learners to evaluate the practical impact of ETEL in real-world educational settings. Gathering qualitative feedback on the system's usability and the perceived accuracy of its readability assessments will guide future iterations of the web portal and API, ensuring that ETEL remains a user-centric tool for the Basque language community.

### *Acknowledgments*

This work has been partially funded by the Vice-Rectorate for Basque, Culture, and Internationalization of the University of the Basque Country (UPV/EHU) and by the CLARIAH-ES infrastructure.

## *References*


Allkivi, K. 2026. Towards interpretable models for language proficiency assessment: Predicting the cefr level of estonian learner texts. *arXiv preprint arXiv:2602.13102*.

Azurmendi, E., X. Arregi, and O. Lopez de Lacalle. 2025. Automatic essay scoring and feedback generation in basque language learning. *arXiv preprint arXiv:2512.08713*.

Bengoetxea, K. and I. Gonzalez-Dios. 2021. Multiaztertest: A multilingual analyzer on multiple levels of language for readability assessment. *arXiv preprint arXiv:2109.04870*.

Benjamin, R. G. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88.

Collins-Thompson, K. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Devlin, J., M.-W. Chang, K. Lee, and K. 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.

Gonzalez-Dios, I., M. J. Aranzabe, and A. Díaz de Ilarraza. 2014. Simple or complex? assessing the readability of basque texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 359–369.

Gunning, R. 1952. *The technique of clear writing*. McGraw-Hill.

Kincaid, J. P., R. P. Fishburne, Jr., R. L. Rogers, and B. S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Lee, J. and S. Vajjala. 2022. A neural pairwise ranking model for readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3296–3303.

McNamara, D. S., A. C. Graesser, P. M. McCarthy, and Z. Cai. 2014. *Automated evaluation of text and discourse with CohMetrix*. Cambridge University Press.

Naous, T., M. J. Ryan, A. Lavrouk, M. Chandra, and W. 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.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rodriguez Rey, S. 2024. Automatic text classification using readability levels in galician and spanish. In *Doctoral Symposium on Natural Language Processing 2024*.

Vajjala, S. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377.

Vajjala, S. and T. Rama. 2018. Experiments with universal cefr classification. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 234–240.