Automated Essay Scoring and Language Certification: Assessing Generalizability, Agreement and Validity for French

Rodrigo Wilkens[◇] and **Rémi Cardon**[♠] and **Vincent Folny**[♡] and **Thomas François**[♠]

[◇] University of Exeter, [♡] France Éducation international, [♠] Cental, IL&C, UCLouvain
[♠] Computer Science and Engineering Department, Universidad Carlos III de Madrid

r.wilkens@exeter.ac.uk, rcardon@inf.uc3m.es
Folny@france-education-international.fr thomas.francois@uclouvain.be

Abstract

In Automated Essay Scoring (AES), benchmarking practices have fostered minimalist evaluation practices, in contrast with the broader-view recommendations of evaluation frameworks, such as the argument-based validation framework (ABV), which argued in favor of a multidimensional assessment of systems, especially in the context of high-stakes language tests. In this paper, we introduce an enhanced and more practical version of the ABV framework, incorporating fairness analysis, correlations with linguistic features, prediction error evaluation, and model agreement compared with human raters. Applying this framework to French AES, we compare 8 model architectures on a corpus of 27k exam essays (2 raters each) and a generalization corpus of 961 essays (at least nine raters each). Our analyses illustrate the benefits of applying the ABV framework to better understand the capabilities and pitfalls of AES models, while also advancing the state-of-the-art for French AES.

1 Introduction

Automated Essay Scoring (AES)[1] aims to produce systems that imitate groups of expert raters by combining principles of educational assessment and Natural Language Processing (NLP). The goal is to provide a valid, cost-efficient and reliable solution for the automatic evaluation of essays (Klebanov and Madnani, 2021). In this paper, we focus on certification tests[2], which are a type of standardized language test defined as "a process by which individuals are certified as having demonstrated some level of knowledge and skill in some domain" (AERA et al., 2014, p. 216). Certification tests are generally high-stake, meaning that outcomes such as "admission, promotion, placement or graduation are directly dependent on test scores" (Shohamy et al., 1996, p. 300) in contrast to low-stake exams (e.g., formative evaluation). In view of the social consequences of high-stake tests, their validity and reliability are critical, as already pointed out by Messick (1990, p. 17).

AES systems tend to be scalable and flexible (Klebanov and Madnani, 2021), and reliable and reproducible (Yan et al., 2020), which are desirable characteristics for real-world utilization. Recent AES literature, especially in NLP, has focused more on models' improvements or new architectures than on the assessment of validity. For ease of comparison, most studies operate within a controlled setting (i.e., the ASAP dataset and a single metric) (Kusuma et al., 2022). However, this does not align with the requirements of high-stakes language tests. Commonly overlooked elements in AES evaluation include the results' generalizability across different corpora, and the relationship between the model and the human raters' agreement. Yet, those elements are required for standardized language certification tests, such as IELTS or TCF.

In language testing, evaluating the tests' validity and reliability has been extensively investigated, using the Argument-Based Validation (ABV) framework (Chapelle et al., 2008), best developed by Williamson et al. (2012). This tradition has had little impact on NLP practices. Aiming to bridge this gap, we perform extensive experiments in the context of one of the major standardized language certification tests for French, the French Knowledge Test (*Test de connaissance du Français* – TCF), and report them within the ABV framework.

This work advances the state-of-the-art for AES, targeting the French language. We compare 8 dif-

[1] Automated Essay Scoring is also called Automated Essay Grading (AEG) (Valenti et al., 2003), Automated Essay Evaluation (AEE) (Shermis, 2015), Automated Writing Evaluation (AWE) (Beigman Klebanov and Madnani, 2020), and Analytic Writing Assessment (AWA) (Rudner et al., 2006).

[2] For a comprehensive list of certification language tests see en.wikipedia.org/wiki/List_of_language_proficiency_tests.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 983–1007, 2026. https://doi.org/10.1162/TACL.a.693
Action Editor: Dilek Hakkani-Tur. Submission batch: 7/2024; Revision batch: 4/2025; Published 6/2026.

ferent model architectures, including hybrid architectures[3], relying on the largest corpus made for French AES. In addition, based on the ABV framework, we report the most extensive evaluation of AES models for French to date, with a strong view on the real-world utilization of our system. We go further in our systematic comparison of various AES approaches, by leveraging a high-quality test corpus made for this specific purpose.

Beyond this specific contribution for French, our main contribution consists in an enhancement of the ABV framework. As it is not easily automatable in its original form, we have identified relevant automatic metrics from NLP and systematically linked them to the framework. Thus, we offer a practical AES evaluation framework for language testing. The focus of our approach is aligned with the recommendations for adding automated scoring to an existing assessment (Madnani and Cahill, 2018). We argue that, by bridging the gap between methodologies focused on model validity and more theory-grounded arguments, such a methodological approach will enable enhancements to automatic evaluation practices for AES.

The outline of the paper is as follows. We first provide an overview of the literature on AES, focusing on models, their limitations for certified exams, and an introduction of the ABV framework (Section 2). We then describe how we approach this framework that will guide our study in Section 3. Next, we describe the models' architectures we compare (Section 4) and the corpora we work with (Section 5). Finally, we present the results of our experiments, structured in relation to our enhanced ABV framework (Section 6), discuss our findings (Section 7), and conclude (Section 8).

2 Approaches for AES

2.1 AES Models

AES has a long tradition of feature engineering methods, as they facilitate studying the relation to the test construct. According to Zesch et al. (2015), features can be classified according to their degree of task dependence, ranging from weak to strong. They showed that task-independent features transfer better between prompts, while the task-dependent ones do not generalize well but perform better. This is especially important in the context of certified exams, where prompts need to be regularly updated. Examples of feature-based works are Burstein and Chodorow (1999) with discourse features, or Zesch et al. (2015), which considers readability features, grammar errors, and task and topical similarity.

AES research then transitioned to deep learning, inspired by the results achieved in other NLP tasks. This led to the development of models independent of linguistic features. Taghipour and Ng (2016) explored CNN and LSTM architectures, replacing linguistic features with word embeddings. They observed that combining results from models trained with different samples and architectures improves performance. Building on this, subsequent studies incorporated task-specific knowledge into the models, e.g. coherence as part of the architecture (Tay et al., 2018) and rating criteria (Liang et al., 2018). Yet, Jin et al. (2018) obtained good results by bringing back linguistic features into the models, opening the avenue for hybrid models.

Motivated by the strong performance of large language models, Rodriguez et al. (2019); Mayfield and Black (2020) explored the use of BERT for AES, obtaining discouraging results. Yang et al. (2020) achieved promising results by incorporating multiple loss functions (regression and ranking) in their R2BERT model. Revisiting hybrid methods, Uto et al. (2020) proposed to concatenate 25 handcrafted essay-level features (length-based, syntactic, word-based, and readability features) to BERT distributed representations. Exploring generative models, Mizumoto and Eguchi (2023) applied GPT-3 to AES, finding a 52% exact agreement with the gold-standard, and demonstrating that the inclusion of linguistic features improves GPT-3's predictions. Finally, Kusuma et al. (2022), in their literature review, found that the best-performing models – reaching 0.801 Quadratic Weighted Kappa (QWK) on the ASAP dataset – are SBLSTMA (Liang et al., 2018) and BERT with handcrafted-features (Uto et al., 2020). Xie et al. (2022) presented a method combining regression and ranking that obtained 0.817 QWK on ASAP.

As regards AES for French, research remains very limited. Early studies employed unsupervised methods: Lemaire and Dessus (2001) used Latent Semantic Analysis (LSA) to compare native students' essays with reference textbooks, whereas AUTO-EVAL (Zaghouani, 2002) extracted and combined features of L1 essays. Later, Parslow

[3] https://gitlab.com/rswilkens/linguistic-features-in-transformers

(2015) trained a Naive Bayes classifier on a very small corpus of 200 FFL (French as a Foreign Language) essays. Ranković et al. (2020) became the first to fine-tune BERT on French data, whereas Wilkens et al. (2023) recently released the largest FFL corpus (6.5k essays) and fine-tuned CamemBERT on it. Muñoz Sánchez et al. (2024) used the same model and corpus, but focused on freezing layers instead of full fine-tuning. Both works report similar F1 scores (0.56 vs 0.57, with a standard deviation of 0.01 in both approaches). None of these studies has sought to go beyond a performance-driven approach to evaluate their models from a real-world perspective.

2.2 Limits for Certification Exams

The need to compare models on reference data has led the majority of AES studies to rely on the same English dataset: ASAP (Kusuma et al., 2022). Only a few studies have departed from this perspective, relying on certification exam datasets: for English, *Cambridge Learner Corpus* (Nicholls, 2003), the *ETS Corpus* (TOEFL11) (Blanchard et al., 2013), with 12,100 essays written by TOEFL candidates, and *OpenCLC* (Corpus, 2017), which contains 1,238 texts following with the CEFR. For other languages, *MERLIN* (Boyd et al., 2014) with 2,290 texts targeted German, Italian and Czech, whereas *COPLE2* (Mendes et al., 2016) and *ASK* (Tenfjord et al., 2006) are for Portuguese with 966 and 1936 texts, respectively. Finally, corpora for French are scarce and suffer from shortcomings: they are either too small (Parslow, 2015) or not available (Ranković et al., 2020), with the exception of TCFLE-8 (Wilkens et al., 2023).

Another limitation in AES studies, particularly sensitive for high-stakes language exams, is their exclusive reliance on the QWK metric for evaluation. Doewes et al. (2023) highlighted shortcomings in QWK, emphasizing that relying solely on it could compromise reliability. Moreover, most AES models adopt an essay prompt-based approach, requiring one model per essay prompt, which is impractical for language testing[4], as prompts are frequently updated (Jiang et al., 2023).

2.3 Framework for High-Stakes Assessments

Some AES implementations are specifically designed for high-stakes assessments. The first system is e-rater (Burstein et al., 1998), developed for the GMAT test, which uses different feature extraction techniques. This system and others that followed – like IntelliMetric (Rudner et al., 2006) – are exploited conjointly with human raters for tasks within tests such as TOEFL (Attali, 2009) or GRE (Bridgeman et al., 2009). One example of an AES system operated without human raters' intervention is IEA for the Pearson Test of English (Pearson, 2019), which relies on LSA.

High-stakes AES requires specific attention to evaluation (Weigle, 2013; Ramineni et al., 2012). Some researchers, concerned with the validity of AES in high-stakes contexts, have contributed to the development of several validity frameworks for automated scoring (Bennett and Bejar, 1998; Clauser et al., 2002; Xi, 2008). Gradually, these reflections moved closer to Kane's argument-based approach to test validation (Kane, 2013), which is rooted in the field of educational measurement. This trend is illustrated by the works of Clauser et al. (2002) Xi (2008) and Williamson et al. (2012), and offers a more comprehensive validation process, encompassing varied evidence sources. More recently, the ABV framework was used by Huawei and Aryadoust (2023) as a means of organizing a systematic review of the field. They identified mixed results when comparing different studies, most of which focused on only a few areas of the framework.

In this work, we draw from one of the most refined presentations of the ABV framework by Williamson et al. (2012), who proposed an operationalization of the summative stage (Kane, 2013) through five areas of emphasis, which they apply to e-rater for illustration. The strength of Williamson et al. (2012)'s framework is that it provides several high-level guidelines criteria, and best practices, for each area. In the rest of this section, we briefly introduce these five areas.

Construct relevance and representation (1) focuses on AES model *explanation* by assessing the adequacy between the goals of the assessment (e.g., evaluating linguistic proficiency) and the automated scoring capabilities of the system. Williamson et al. (2012) propose four steps to evaluate this fit: construct evaluation, task design, scoring rubric, and reporting goals.

Association between typical scoring method and human scores (2) focuses on the *evaluation* and addresses the limitations of using human scores

[4] Jin et al. (2018) identified a performance drop of 0.225 points of QWK when comparing prompt-agnostic and prompt-dependent scenarios.

as the "gold standard" when evaluating AES systems with automatic metrics (e.g., QWK). The authors suggest a set of additional analyses along with some reference thresholds: first, assess the quality of the human scores; then, report the QWK between human and automated scores using a threshold at which half of the variance of human scores is accounted for by the system (0.70 in their case); measure the difference between system-human and human-human agreements (computed with QWK or correlation) and evaluate that difference against a minimum threshold (-0.10 in their case); report the standardized mean score difference between human and automated scores (standardized on the distribution of human scores), with a maximum threshold of 0.15; assess the threshold required to initiate human adjudication (i.e. using an extra rater); analyze the type of response characteristics that invalidate automatic scoring; and finally, assess the impact of the system on individual tasks or aggregated reported scores by simulating substitutions between automated and human scoring.

Association with independent measures (3) suggests leveraging independent measures, such as scores from another section of the test or external measures of the construct (i.e., *extrapolation*). Specifically, automated scores should have the same relation as human scores on other sections of the test, and align with external measures of the same construct in the same way as human scores. No specific metric is recommended, but the authors acknowledge that divergences hardly occur in AES.

Generalizability of scores (4) questions the applicability of a single system to different tasks by analyzing facets (e.g., task types, prompts, genres, or populations of test-takers). It corresponds to the *generalization* area of the framework. It includes analyses with other scores than the gold standard, such as a single rater's score or an alternative aggregation of two human raters' scores. No specific metric is recommended in the original framework.

Score use and consequences (5) is specific to the implementation of an AES system in a high-stakes environment (i.e., *utilization*). The authors suggest analyzing the aggregated reported scores in depth to measure the impact on the final decision errors compared to human scores. Again, there is no specific recommended approach. However, the recommendation is to examine how automated scoring affects the accuracy of decisions, the claims and disclosures necessary for score users to ensure appropriate score interpretation, and the impact of automation on the test environment. In addition, the effects of automation should not be solely analyzed globally, but also across different subgroups to assess the fairness of the AES system.

3 Evaluation Framework Methodology

To assess our models, we extend the ABV framework proposed by Williamson et al. (2012), designing our evaluation around the five areas they outlined (introduced in Section 3.1). Each area is addressed through one or more automated scores (Section 3.2) that either follow or extend Williamson et al. (2012)'s framework. This section aims to pave the way for an automated assessment of AES systems based on the argument-based validity (ABV) framework, to improve the evaluation protocols of future AES systems. The mapping between specific metrics proposed in this work and the five areas from the ABV framework is described in the rest of this section and is summarized in Table 3.2.

3.1 Indicators for the ABV Framework

In this section, we report how we systematically expand on the ABV framework from Williamson et al. (2012), outlining our quantitative evaluation.

The *(1) construct relevance and representation*, which measures the characteristics of the construct and the rubrics, naturally aligns with linguistic skills. While these skills can be precisely mapped to characteristics linked to the rubrics, this process becomes cumbersome in a holistic evaluation, as it requires annotating the corpus with these characteristics. Consequently, we opt for a more straightforward approach to approximate those, using automatically extracted linguistic features. We study (a) the correlation of the linguistic features in the following ways: correlation between linguistic features and CEFR levels in the scoring rubric assessment, correlation between linguistic features and CEFR levels across the tasks (in our case, three) in the construct assessment, and correlation between linguistic features and CEFR levels for each prompt in the task design assessment. In language testing, the number of prompts is generally high and the number of essays per prompt low. Thus, to prevent statistical bias,we only retain prompts associated with more than 100 essays and compute the mean and standard deviation of the correlations.

The *(2) association with typical scoring method* relates to model evaluation and is the most often

Emphasis Area	Automated Scores
(1) Construct Relevance and Representation (explanation)	*(a) linguistic features correlation with human rater* by rubric, tasks and prompt
(2) Association with Human Scores (evaluation)	*(b) ranking* (Spearman's correlation) and *(c) label identification* (MSE, EA, AA and F1), and *(d) agreement* (Cohen's Kappa and QWK)
(3) Association with Independent Measures (extrapolation)	out of scope for this study, as it requires additional scores
(4) Generalizability of Scores (generalization)	*(b) ranking* and *(c) label identification*, and *(d) agreement* on the Generalization Corpus
(5) Score Use and Consequences (utilization)	*(e) fairness* (OSA, OSD, CSD) and *(f) failure identification* (ECE and the relation between models' correctness and human agreement)

Table 1: Alignment between framework and automated scores

reported area of the ABV framework (Williamson et al., 2012). In this work, we focus on three objectives: (b) ranking, (c) label identification, and (d) agreement, all of which compare the system's predictions to the gold standard CEFR levels. The first one examines how the AES model ranks the essays, while the second assesses how it assigns labels (e.g., CEFR levels) to essays. The last one evaluates the extent to which the model's predictions align with raters' judgments, correcting for agreement that might arise by chance. Williamson et al. (2012) recommend using QWK. In addition, examples from the literature point out the need for other common metrics from classification and regression tasks (e.g., Klebanov and Madnani (2021)). These metrics align with recent AES practices with machine learning. On top of this analysis on our entire corpus, we investigate how agreement between human raters impacts model performance. To this end, we analyze subsets of our corpus (see Section 5.1). We propose three categories for sampling the subsets: *full agreement* (human raters indicate the same score), *close agreement* (difference of 1 level), and *weak agreement* (difference of 2 levels).

For *(3) association with independent measures*, independent scores are necessary, as mentioned, to assess extrapolation. Since such additional scores are rarely available in AES research and published datasets and were not accessible for the TCF, this aspect falls outside the scope of this paper.

The *(4) generalizability of scores* area assesses the applicability of the scoring system across different datasets. It evaluates performance under various test conditions (e.g., different tasks or different populations of test takers). For the generalization analysis, we assess the model's performance (i.e., (b) ranking abilities and (c) label identification, and (d) agreement) on a corpus with unseen essays sampled to be representative of the target testing scenario, reflecting the model's application scenario. In addition, essays in this corpus have undergone a much more rigorous evaluation process, ensuring higher quality (see Section 5.2 for details).

The *(5) score use and consequences* targets high-stakes environments and the potential for decision errors. To evaluate this area, we rely on two measures. First, we study model (e) fairness to identify biases regarding criteria unrelated to the scoring task (e.g., gender or native language). We also assess (f) failure identification, i.e., the possibility of identifying incorrect predictions of the model.

At this point, we have established six analysis categories corresponding to four out of the five areas of the ABV framework, with some categories applying to multiple areas. However, to produce an automatable assessment framework, these categories must be operationalized with concrete metrics, which we do in the next section.

3.2 Construct Validity Metrics

Building on the six categories of metrics defined in Section 3.1, this section introduces the concrete measures used within each category. This aims to ensure a reproducible evaluation protocol while preserving the interpretability of the original framework proposed by Williamson et al. (2012).

Most AES systems are automatically evaluated using direct comparison analysis, which assesses the relation between predictions and ground truth.

There is a tendency in the literature to report a single metric, QWK. As this may prevent capturing the full complexity of the models' performance, we follow Kusuma et al. (2022)'s recommendation of using a wider range of evaluation metrics, aiming to minimize potential evaluation errors. For the (b) ranking metrics, we select the Rank Spearman's correlation (ρ), which evaluates only the order of predictions without considering the label. For the (c) label identification metric, we follow Klebanov and Madnani (2021), measuring the model's Exact Agreement (EA) for accuracy, Adjacent Agreement or Adjacent Accuracy (AA) for distinguishing between minor and severe errors, Mean Squared Error (MSE) for measuring the distance to the expected scores, and macro F_1 ($F1$) for assessing the balance between labels. Although other metrics exist for this task, we chose these as they are most cited as best practices and studied in educational assessment or AES (Gwet, 2014; Lottridge et al., 2023; McCaffrey et al., 2022; Stemler and Tsai, 2008; Von Eye and Mun, 2014; Yan and Bridgeman, 2020).

The (d) agreement between the human and model labels relies on Cohen's Kappa (κ) and QWK scores in this work. κ evaluates the agreement between predicted and actual labels while accounting for chance agreement, while QWK also considers the ordinal nature of predictions when computing the agreement. Note that κ and QWK could also be considered metrics of label identification and label ranking, respectively.

The (f) fairness analysis aims to identify potential biases regarding candidates' characteristics that are unrelated to the task (e.g., gender or prompt). The models' performance is examined through the lens of those characteristics. Our fairness analysis follows the method proposed by Loukina et al. (2019), evaluating three key metrics: overall score accuracy (OSA) assesses whether the model maintains equal accuracy for each group by comparing differences in squared error between human and automated scores; overall score difference (OSD) examines whether the model's predictions systematically differ from human scores for members of a certain group; and conditional score difference (CSD) verifies whether the model fairly assigns scores to candidates from different groups.

The (a) linguistic correlation analysis aims to measure the degree of correlation (using Spearman's ρ) between scores and linguistic features related to the construct. These correlations help to assess how the models' predictions are aligned with the relevant underlying linguistic characteristics for language assessment. As correlation estimates can be computed at the rubric, task and prompt level, the number of data points may be insufficient to yield robust estimates (especially at the prompt level), as small sample sizes can introduce high variance. To mitigate this, we exclude cases with fewer than 100 essays and those where fewer than 5 of the 6 levels were represented.

Finally, the (f) failure identification analysis aims to evaluate the model's trustworthiness. To this end, we assess the models' calibration (i.e. the relation between the probability of a predicted level and its correctness) by calculating the Expected Calibration Error (ECE) (Guo et al., 2017). Moreover, we analyze the relation between models' correctness and human agreement, dividing the corpus into cases where all models are correct, all models are wrong, or mixed. We report the average human agreement in each group.

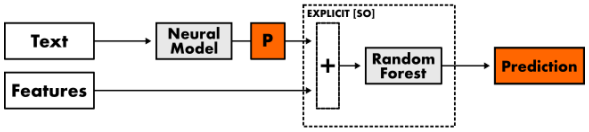
4 Explored Models

For a comprehensive comparison of AES architectures, we follow the current preference for hybrid approaches in AES by combining linguistic features and deep learning. The models explored in this work were previously validated in the context of French readability assessment (a reception task) (Wilkens et al., 2024). Here, we extend their application to AES (a production task). While the tasks differ, the architectures align with state-of-the-art AES architectures.[5]

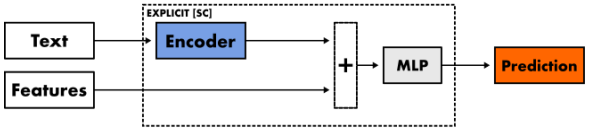
Baseline: we use a transformer encoder (*TR*) based on RoBERTa, more specifically CamemBERT (Martin et al., 2020), as a non-hybrid baseline for measuring the contribution of the features. This model was previously explored for French AES by Wilkens et al. (2023) and Muñoz Sánchez et al. (2024) using the TCFLE-8 corpus.

Direct (or explicit) combination: We explore two approaches. The first one leverages ensemble methods by using *soft-labeling (SL)* and *hard-labeling (HL)*. In both methods, the TR model is first fine-tuned, then its predictions are combined with features and used to train a Random Forest classifier (see Figure 1a). With SL, the softmax output is used in the concatenation (see Figure 1a), whereas only the final prediction is used with HL.

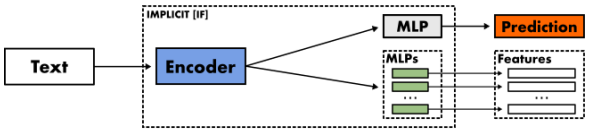
[5]Models' hyperparameters are discussed in Appendix A.



(a) Soft-labeling



(b) Simple Concatenation



(c) Implicit Feature

Figure 1: The hybrid architectures

The second strategy which is similar to the one by Uto et al. (2020), the most common hybrid approach, consists of feeding a multilayer perceptron (MLP) with the concatenation of the TR's output (i.e. the CLS) and the features. We name this approach *Simple Concatenation (SC)* (see Figure 1b). Additionally, we can improve the network capabilities by adding MLPs at different levels of the architecture. An MLP between the features and the concatenation is here named *Concatenate MLP (CM)*, and a CM with an additional MLP after the encoder is here named *Concatenate 2xMLP (C2)*.

Indirect (or implicit) integration: Language features can also be imprinted on the network through the use of auxiliary tasks, following a multi-task approach. The first implicit architecture we explore, referred to as Implicit Features *(IF)*, learns each linguistic feature as an independent regression task using an MLP (see Figure 1c).[6] A variant of IF, named *Implicit Feature Vector (IV)*, groups all features into a single output vector.

4.1 Linguistic Features

To compute the linguistic features needed for our analyses, we used the FABRA toolkit (Wilkens et al., 2022), originally designed for readability assessment but also including features relevant to AES. FABRA identifies over 400 linguistic phenomena that are parameterized into features using 18 statistical aggregators, resulting in a total of 5,986 features[7]). After annotation, we computed Spearman's, Pearson's and biserial correlations as well as Mutual Information Gain measures for each feature, using the essay's level as the target. This information was analyzed by an expert FFL teacher, highly familiar with corpus annotation and feature engineering. The best features were then selected according to their relationship with essay level, pedagogical relevance, and informativeness.

5 Corpus

Corpora for language testing research are scarce, especially for French (Wilkens et al., 2023). In this work, we leverage access to an extensive collection of professionally-rated texts to investigate AES for French as a foreign language (FFL) certification. The data we use are collections of essays coming from one of the main French certification tests, the French knowledge test (TCF), run by *France Education International*[8] (FEI). In 2024, candidates who sat for the computer-based TCF came from 88 countries and 350 test centers, representing 207 languages. A major challenge in automating a certification test with such global reach is test-taker heterogeneity. Some learned French as a foreign language, others as the language of schooling, and still others as a second or even a first language. TCF is available in several versions: a version targets the general public, while others target different aims such as naturalization and residence, and immigration to Canada. To some extent, the versions embody the diversity of profiles. As regards the evaluation of candidates' proficiency within the TCF, written performances are always independently double-rated, with a third rater if there is a discrepancy of more than one CEFR level. Research shows that for an exam with three tasks, double rating ensures sufficient reliability (Bouwer et al., 2015).

From TCF essays, two datasets were compiled for two different purposes: the core corpus and the generalization corpus. This section introduces both of them. In both datasets, there was no control for the task type and prompt that were used, which means that a variety of types and prompts are represented (based on random selection). This is to ensure that the results are not skewed towards a specific setting with regard to those aspects.

[6] we apply a weight of 0.5 for the loss associated with the target task and 0.5 for the sum of all the feature regression losses.

[7] The entire list of variables is available at `cental.uclouvain.be/fabra/docs_expert.html`

[8] `www.france-education-international.fr`

5.1 Core Corpus

To build up our core corpus, we started from a raw dataset of 450,000 TCF essays provided by FEI [9]. This dataset underwent multiple levels of cleaning. We followed the approach of Wilkens et al. (2023) to decide which outliers to remove and how to set the essay gold scores. This score is determined based on a combination of the candidate's overall performance across the three written tasks of the TCF (narrative, descriptive, and argumentative) – i.e. the candidate's CEFR level – assigned by the raters. More precisely, for a given essay, the assigned score is the candidate's CEFR level when at least one of the raters also assigned that level to the essay. Alternatively, if both raters agreed on the essay's score, we duly assigned this level to the essay. Any essay that did not meet either of these conditions has been removed. It is important to emphasize that this French corpus is labeled according to a candidate-based evaluation, contrary to most available AES corpora. Therefore, the level is assigned to the candidate by combining the scores of each essay. The essays do not have a single CEFR score; instead, they have two (in some cases three), assigned by different raters.

Next, we extracted a stratified random sample from the corpus, balancing the distributions of CEFR scores, adjusted scores[10], and the language of use (the language most frequently used by the learner in daily life). For sampling purposes, we considered only the most frequent candidate languages in the dataset (English, Arabic, Spanish, French[11], Kabyle, Portuguese, and Russian), while the others were grouped into a category named *Other*. In contrast to TCFLE-8, restricted to 8 languages of use, our dataset contains over 164 languages, providing a more comprehensive and representative depiction of the authentic situation of AES for French high-stakes testing. This resulted in a dataset of 27,683 essays – henceforth the **core corpus** – with a rather balanced distribution over the 6 CEFR levels and 852 distinct prompts[12]. In this corpus, 14,141 essays were given the same level by both raters (hereafter *full agreement*), 12,750 had a difference of one level (*close*), and 791 had a difference of two levels (*weak*).

5.2 Generalization Corpus

FEI built an additional corpus of 961 finely calibrated essays – henceforth the **generalization corpus**. It has been shown that using a carefully crafted corpus helps as a source of validity evidence to the evaluation of AES systems (Loukina et al., 2018). This generalization corpus comprises essays sampled to represent as much as possible the distribution of the real-world scoring scenario of the TCF. To maximize the reliability and accuracy of the gold scores of this corpus, each essay was rated by a large number of raters: between 9 and 55 raters (depending on inter-rater agreement). Such a process was needed because of deviations in human ratings (Klebanov and Madnani, 2021).

6 Results

We experimented with 8 AES architectures (Section 4), all trained using the core corpus (Section 5). For each architecture, we trained models using stratified 10-fold cross-validation[13]. In this section, we evaluate these 8 models based on the revised framework and all metrics described in Section 3.2. We thus present the correlation results (Section 6.1), the direct comparison analysis including ranking, label identification and agreement (Section 6.2), the fairness study (Section 6.3), and the error identification study (Section 6.4).

6.1 Linguistic Correlation Analysis

We start by reporting interesting findings regarding the behavior of the 48 features selected as a result of the process described in Section 4.1.[14] Among this selection, "occurrence of errors" shows the highest correlation with the gold learner proficiency levels ($\rho = -0.79$), followed by lexical features capturing word diversity (0.64), the proportion of A1 words

[9] Their availability for research has been granted through an agreement with FEI, including aspects relative to GDPR compliance. Unfortunately, the terms of use prevent us from publicly releasing the corpora used in this study.

[10] An adjusted version of the essay scores were produced through "Many-facet Rasch measurement", a psychometric approach that establishes a coherent framework for drawing reliable, valid, and fair inferences from rater-mediated assessments, thus answering the problem of fallible human ratings (Eckes, 2009).

[11] Candidates may indicate that their language of habitual use is French. This might include individuals living in French-speaking countries or native speakers who need to attest their proficiency in French.

[12] A complete description of the size of the core corpus can be seen in Table 5 in the appendix.

[13] 10% of the dataset is used for validation and 10% for test. Consequently, 10 models were trained for each architecture with an 80/10/10 dataset split.

[14] The list of features, with their correlation, is shown in Table 7, Appendix D.

in the text (-0.61) and the average orthographic Levenshtein Distance to the 20 closest neighbors (Yarkoni et al., 2008) (0.57). The traditional word length feature – computed as the number of syllables per word – reaches 0.60, while sentence length has a lower correlation (0.43). Among the syntactic-based features, sentence depth is the most correlated one (0.52). The relatively strong correlations confirm that these linguistic features are well-suited to capture written proficiency, aligning with the goals of the assessment. Although the remaining selected features have a correlation coefficient below 0.5, they were retained in our models in order to diversify the information.

Next, to further study the associations between linguistic features and gold proficiency levels, we examine how these correlations evolve when the proficiency levels are those of a single rater (human or model). For each rater and each feature, we compute the Δ between the correlation on the gold and the correlation of the given rater. Table 2 (column Δ **Corr**) shows the results of this analysis.

On the core corpus, we observe that all raters have a minimal delta between their correlations. The human raters show an average difference of 0.01 for all features, while the models have values of about -0.01 (standard deviation < 0.001 in all cases). This indicates a high level of stability in relation to features among all raters. The models tend to be slightly more aligned with the features than the human raters. We repeat this procedure at both the task and prompt level. We observe a similar Δ between correlations across the 3 tasks. Regarding prompts, we exclude all cases with fewer than 100 essays (see Section 3.1) and focus on the remaining 54 prompts, where we observe the same general trend as in the larger corpus. Human raters as well as the transformer (TR), the *simple concatenation* (SC) and the *concatenate MLP* (CM) models produce similar results, while the *concatenate MLPx2* (C2), and both implicit models (IF and IV) show a slight reduction of the Δ. The *hard-labeling* architecture (HL) exhibits the largest difference (-0.03). These results indicate that task and prompt effects are independent in the models.

Focusing on samples grouped by raters' agreement (*full*, *close*, and *weak*), the average differences between humans and models remain comparable. Additionally, there is a clear increase in the mean differences as agreement decreases.

6.2 Models Performance

6.2.1 Standard evaluation on the core corpus

The results of the Direct Comparison Analysis are shown in Table 2, where we can observe a remarkable stability. Only small performance differences are associated to the architectures, which confirms the findings of Rodriguez et al. (2019) and Mayfield and Black (2020) about BERT efficiency. When considering the entire core corpus, the automated models fall behind the human raters[15] as regards κ, QWK, EA, and $F1$. A closer examination of the models' performance and raters' agreement reveals a strong performance of human raters. This result is partly explained by the fact that their scores contributed to the creation of the gold standard. Interestingly, this bias disappears when comparing cases of full agreement with those of weak agreement, where human raters actually perform worse than the models in ρ, QWK, MSE and AA (see Table 2). Moreover, comparing the metrics that penalize all errors equally (e.g., $F1$ and κ) and those that penalize errors involving adjacent levels less (e.g., QWK, AA, ρ), we see that most model errors are off by one level (e.g., A1 or B1 instead of A2).

Focusing on the architecture of the models, we observe in Table 2 that Transformer (*TR*) is competitive with the hybrid models. Although the top-performing models vary depending on the combination of corpus sample and the evaluation metric, the *soft-labeling* strategy (*SL*) tends to perform the best on the core corpus as well as on samples with full and close agreement. *Simple concatenation* (*SC*) is slightly better on the weak agreement sample.

As regards the explicit hybrid models, they do not differ statistically from each other on the core corpus, even though *SC* seems to perform slightly better. On the contrary, the implicit models differ statistically from each other (*IV* achieves better performance), and the ensemble models (*SL* and *HL*) too. If we only consider the top model from each architecture family, the *soft-labeling* strategy (*SL*) outperforms the other architectures, followed by *simple concatenation* (*SC*) and Transformer (*TR*)[16].

A more fine-grained inspection of the results re-

[15] We compute the agreement between the raters, using Cohen's kappa and QWK. In this corpus, a large number of raters participated in the annotation, but for the sake of simplicity, we do not distinguish individuals and compare "rater 1" and "rater 2" for each text, obtaining an agreement of 0.4 for κ and 0.83 for QWK. This agreement is similar to those observed in the literature (Taghipour and Ng, 2016).

[16] *SC* and *TR* do not statistically differ from each other.

Corpus	Model	Direct Comparison							Δ
		ρ	κ	QWK	MSE	EA	AA	$F1$	**Corr**
Core Corpus	TR	0.90	0.48	0.89	0.50	0.57	0.98	0.57	-0.02
	CM	0.90	0.48	0.89	0.49	0.57	0.98	0.57	-0.02
	C2	0.90	0.48	0.89	0.49	0.57	0.98	0.57	-0.02
	SC	0.90	0.44	0.88	0.55	0.54	0.97	0.54	-0.02
	IF	0.90	0.48	0.89	0.48	0.58	0.98	0.57	-0.02
	IV	0.90	0.48	0.89	0.49	0.57	0.98	0.57	-0.01
	SL	0.90	0.50	0.89	0.44	0.60	0.99	0.59	-0.03
	HL	0.90	0.47	0.88	0.48	0.57	0.98	0.56	-0.02
	Raters	0.91	0.70	0.91	0.40	0.75	0.98	0.75	0.01
Full agreement	TR	0.92	0.55	0.91	0.40	0.63	0.99	0.64	-0.01
	SC	0.92	0.51	0.90	0.45	0.60	0.98	0.60	-0.01
	IV	0.92	0.55	0.91	0.39	0.63	0.99	0.64	-0.01
	SL	0.92	0.57	0.91	0.36	0.66	0.99	0.65	-0.02
	Raters	1.00	1.00	1.00	0.00	1.00	1.00	1.00	-0.00
Close agreement	TR	0.89	0.41	0.87	0.59	0.52	0.97	0.51	-0.02
	SC	0.89	0.37	0.86	0.65	0.48	0.96	0.47	-0.02
	IV	0.89	0.41	0.87	0.57	0.52	0.97	0.50	-0.02
	SL	0.88	0.40	0.86	0.56	0.52	0.98	0.49	-0.02
	Raters	0.87	0.40	0.86	0.61	0.51	0.99	0.50	0.01
Weak agreement	TR	0.84	0.23	0.82	0.87	0.36	0.92	0.36	-0.05
	SC	0.83	0.24	0.82	0.91	0.37	0.91	0.37	-0.06
	IV	0.83	0.24	0.82	0.85	0.37	0.93	0.35	-0.05
	SL	0.83	0.21	0.81	0.84	0.35	0.94	0.33	-0.06
	Raters	0.67	0.33	0.66	1.88	0.45	0.57	0.44	0.06

Table 2: Results of the direct comparison and linguistic correlation analysis in the core corpus, including its splits by human raters agreement (the results for all the models are presented in Table 8).

veals that the models' scores tend to decrease as the CEFR level increases (see Table 11), in contrast with human raters. While the best-performing model at a given level varies, some consistent patterns emerge. All the models (except HL) perform best at the A1 and A2 levels, which contain texts with highly typical lexico-grammatical features (short sentences, simple words, etc.). From level B1 onwards, however, there is a general deterioration, which becomes more important and culminates at the C2 level for some models (C2, SC, IF, and especially SL and HL). A possible explanation for this performance drop at the C levels is that skilled written productions can be characterized by a wider set of linguistic properties (e.g., coherence and cohesion, implicitness, phraseology, etc.). It has been acknowledged in the learner corpus literature (Forsberg Lundell and Lindqvist, 2014; De Clercq and Housen, 2017) that complexity features generally have a weaker discriminative power at C levels. Another analysis consists of comparing the performance of the transformer, which relies solely on the information contained in the embeddings, with that of the hybrid models that also integrate features. It is interesting to note that TR does not exhibit any performance deterioration at the C2 level, whereas the models that rely most heavily on the variables (in particular SL and HL) suffer the most. Conversely, these same models seem to benefit the most from the effect of the variables at the B2 and C1 levels. This effect is weaker or even absent for the explicit models, perhaps because the presence of MLP layers mitigates it.

Looking at the classification errors of the best-performing model, the *soft-labelling* strategy (SL), we can see that it tends to assign a higher CEFR level when the expected level is B2 or below, while it assigns a lower level when the expected level is C1 or C2. Moreover, ~56% of the C2 essays are misclassified as C1, and ~40% are correctly

classified as C2. This confirms the weakness of the model in identifying essays of level C2 compared to the gold.

6.2.2 Evaluation on the generalization corpus

Evaluating our models on the generalization corpus brings more insights. The results of the direct comparison metrics are presented in Table 3. We only report on the best architecture from each family. The most prominent result is the performance gain for the generalization corpus over the core corpus. Specifically, this difference tends to get smaller when only the full agreement sample is considered. Unlike the results on the core corpus, the human raters' performance decreases. This is explained by the already mentioned fact that the gold scores for the core corpus are based on the decisions of only two raters who are also used in our analysis, resulting in symmetric results when comparing each rater with the gold standard. On the other hand, the reference score of an essay is determined by a large pool of raters in the generalization corpus, thereby reducing the variation introduced by individual raters. Consequently, for comparison purposes, we reported raters' performance compared with the gold scores (Rater 1 and Rater 2) considering only two randomly selected raters out of the larger pool of raters for each essay. Comparing the performance of human and automated raters, we see that the models' performance is equal or superior to the humans' across all measures. This points out the excellent generalization capacity of the models and the critical need for such evaluation corpora to properly assess the performance of AES systems before they are deployed in production. Regarding model performance, *SL* shows superior results across all metrics again.

Regarding the models used to rate the essays of the generalization corpus, it is important to note that 10 different models trained with different samples from the core corpus were used (each model is trained on one of the 10 repetitions from the cross-validation procedure). However, it is also important to bear in mind that a common approach in AES consists in combining models trained on different samples (i.e., as proposed by Taghipour and Ng (2016)). Therefore, we also evaluate ensemble models using the majority vote for prediction.[17] This means that, for each model, we predict the essay levels in the generalization corpus. In Table 3, we indicate the results based on this approach with the *vote* suffix. The voting approach mostly improved the models. However, the *SL* architecture keeps its top position, even with the small improvement caused by the voting approach. This confirms the robustness of this architecture. Interestingly, the difference between *TR* and TR_{vote} is negligible. Despite the slightly superior performance of the voting model in this study, its use implies a significantly higher computational cost. Thus, our results point to the use of SL without voting.

6.2.3 Analyses of the best model

We conducted a series of more in-depth analyses of the best-performing model, namely the *soft-labeling* approach. As previously, we computed the model's performance by CEFR level, but on the generalization corpus, which offers substantially more reliable assessments. Comparing the distribution of SL on the core corpus and the generalization corpus (line SL-GC in Table 11), we see that results on A levels surprisingly drop, and dramatically increase for all other levels, including C2 (in contrast with previous findings on the core corpus). These results call into question our previous hypothesis regarding the limited discriminative power of our variables at the C level. It could be instead that human rating of C2-level essays may be especially difficult to perform consistently. A plausible contributing factor of this seems to be the underspecified nature of the CEFR descriptors at the C2 level[18].

We also studied the distribution of language errors in the essays and its relation to the prediction errors of the *soft-labeling* model (i.e., residuals). For this, we manually analyzed errors made by the candidates in a sample of essays. We subsampled 308 essays in the test corpus, using stratified sampling based on two criteria: the CEFR levels of the texts and values of the residuals. In each essay, we manually classified errors into three categories: lexical, (e.g., spelling mistakes, extra or missing diacritics, non-existing words), morphological (e.g., agreement or conjugation errors) and syntactic. (e.g., extra or missing punctuation/words, word order issues). For each level, we then com-

[17] Our aim is to assess an improvement in the combination of models, rather than to identify the best way of combining them. For a discussion of different methods of combining AES models, see Uto et al. (2023).

[18] An example of this underspecification can be directly found in the Companion volume of the CEFR (Council of Europe, 2020, 41): "For example, at C2 the entry is sometimes: No descriptor available: see C1".

Model	ρ	κ	*QWK*	*MSE*	*EA*	*AA*	*F*1
TR	0.93	0.56	0.91	0.42	0.63	0.99	0.63
SC	0.93	0.56	0.91	0.41	0.64	0.99	0.63
IV	0.93	0.56	0.91	0.43	0.64	0.99	0.63
SL	0.93	0.60	0.92	0.37	0.67	0.99	0.66
Rater1	0.92	0.51	0.88	0.58	0.59	0.96	0.51
Rater2	0.92	0.58	0.89	0.51	0.65	0.96	0.56
TR_{vote}	0.93	0.56	0.91	0.41	0.64	0.99	0.63
SC_{vote}	0.94	0.60	0.92	0.37	0.67	0.99	0.66
IV_{vote}	0.93	0.57	0.91	0.42	0.65	0.99	0.64
SL_{vote}	0.93	0.61	0.92	0.35	0.68	0.99	0.67

Table 3: Results of the direct comparison analysis in the generalization corpus.

pute the Spearman's correlation between the number of errors of each category (normalized over document length) and the model's residuals.

Table 9a displays the proportion of each type of errors across the CEFR levels, indicating a clear decline in errors as proficiency increases. For the correlations between residuals and type of errors (see Table 9b), the only significant ones can be found at low proficiency levels (A1 and A2) and C2 for the lexical category and at C2 for the syntactic category. Those suggest that the model may strongly rely on the presence of lexical errors to identify the A levels (as the model tends to assign a higher level to A-level texts with few errors), and on the absence of lexical errors to identify the C2 level (as C2 texts having more errors are more poorly classified). The text examples in Table 10 (Appendix G) illustrate this heavy reliance on errors: texts classified as B1 by humans based on fluency and syntax are penalized as A1 due to a large amount of lexical errors. We deem this phenomenon worthy of further investigation and leave it as future work, as it would require annotating a larger volume of data.

6.3 Fairness Evaluation Analysis

The fairness analysis shows no bias towards language of use, gender, or task with any of the three metrics. When sampling the core corpus by the degree of agreement between the raters, no gender bias appears neither. Regarding task and language, a small bias can be observed on the weak agreement sample. The bias on language of use amounts to 0.20 of R^2 for overall score accuracy (OSA) for the *C2* model, 0.15 for the overall score difference (OSD) for *CM*, and 0.11 for the conditional score difference (CSD) for *C2*. For task bias in specific models, small biases were identified by the OSA metric (0.1 of R^2 in *HL*) or the OSD metric (0.13 (*IF*), 0.11 (*IV*) and 0.12 (*SL*)). However, this bias is only identified in one out of ten models trained using the cross-validation. In addition, as a proxy for test-taker background, we analyze the performance of models on each of the TCF variants (e.g., Canada and French citizenship) and found no substantial differences. Finally and more importantly, no bias has been identified on the generalization corpus.

The above quantitative fairness analysis was based on classic factors (e.g., gender, language and task). We also investigated complementary background-related factors, which can also shape models' outcomes. We looked at essays with severe classification errors (examples available at Table 10, Appendix G). The 1st row displays an essay with several misspellings (e.g., *mont sanc* instead of *mon sac*, and *cate d'idantite* instead of *carte d'identité*). Indeed, these characteristics are typical and expected of A1-level candidates, but this profile does not reflect the real CEFR level of this candidate. The 2nd row illustrates a case where a systematic spelling pattern (substituting *e* with *è*) yields a text easily understood by a native speaker but challenging for the system. This may have been the result, for example, of a keyboard usability issue or even an eye condition that was not previously identified during exam registration. This shows the value of an evaluation that combines different criteria, as well as the benefit of not relying on a single automated evaluation system.

6.4 Failure Identification Analysis

To explore the relationship between the models' accuracy and confidence with Expected Calibration

Error, we focus on the *vote* models by comparing the distribution of model predictions at each proficiency level to the expected essay levels. We observed the following measurement errors: 0.19 (0.001 of standard deviation) for human raters 0.15 (0.002) for SL_{vote}, and 0.08 (0.004) for SC_{vote}, IV_{vote} and TR_{vote}. These results indicate that the models are closer to the gold standard than the human raters. However, this should not be interpreted as the models achieving superhuman performance.

Taking a deeper look at failure identification, we inspected a possible association between the distribution of the predictions of the 10 models and the raters percentage of agreement.

When considering the average percentage of exact agreement between raters, we observed large differences between cases for which all four models are *correct* (78% of exact agreement), are *wrong* (63%), and 66% for the *mixed* group. Distinguishing correct and wrong cases within the *mixed* set, we noticed that *TR*, *SC* and *SL* tend to be correct when the raters have a higher agreement (68% for $\text{mixed}_{correct}$ v. 64% for mixed_{wrong}) while *IV* shows the opposite trend (64% for $\text{mixed}_{correct}$ v. 68% for mixed_{wrong}).[19]

It is also worth noting that 94% and 91% of essays in the *wrong* and the *correct* groups respectively were rated by the small rater pool (9 raters), while 87% of essays in the *mixed* group were rated by 55 raters. This shows a tendency of models to disagree on the same cases as human raters.

7 Discussion

Our results show that the systems we investigated are valid and competitive with humans. Nonetheless, they must be understood in the context of a broader evaluation framework: the discussion of our findings is therefore structured according to the five areas of Williamson et al. (2012), using the mapping proposed in Section 3.

The results show minimal differences in correlations between human and system raters (see Section 6.1), confirming a good *(1) construct relevance and representation*. This indicates that both types of raters similarly interpret and apply the rubric based on linguistic skills. The small average difference reflects this consistency. The slightly better alignment of models with linguistic features suggests humans might rely on criteria different from those encoded in the models. Moreover, the differences in correlation for cases of weak agreement might indicate a complementarity between human and automatic ratings.

For *(2) association with typical scoring method (human scores)*, performance across architectures, including improvements from voting, demonstrates robustness from AES in ranking and identifying the essay levels. Automated systems outperform human raters across all metrics, indicating a potential for consistent and accurate assessments. The results of linguistic correlation in the samples with varying agreement levels illustrate the impact of human agreement on the scoring criteria. High consistency in full and close agreement samples suggests precise rubric adherence, while on challenging essays (weak agreement sample), the models remain reliant on linguistic features, whereas it seems that human raters also use other information. This points out a likely improvement when combining human and automatic raters. Despite these promising results, we must be cautious in interpreting them. Our analysis suggests a relationship between model errors and the amount of lexical errors made by candidates in an essay.

Regarding *(4) generalizability of scores*, AES models demonstrate stable generalization. While human raters show a performance drop between the core and generalization corpora, AES models improve slightly. Human performance shifted due to the higher number of human raters involved in creating the generalization corpus, thus obtaining a more reliable scoring (in the core corpus, each rater typically contributes 50% of the score decision, while in the generalization corpus a rater may impact up to 10% of the final score). In addition, we notice that the models' performances are underestimated in the core corpus. This is due to the limited number of raters (i.e., only two) which our findings show is insufficient. This goes in the same direction as studies in the language assessment field (Bouwer et al., 2015; Huang and Whipple, 2023) and raises questions about the performance estimation of several AES models trained on corpora where a single rater assigned the score. In our study, this impact is most noticeable in the C2 level, which challenges the human raters. Such findings spotlight the need for broader use of well-calibrated generalization corpora in the field. Although this comes at a high processing cost, the *vote* results show it is possible to improve the model's performance through

[19] Average agreement percentage by the 3 groups and the model correctness is shown in Table 6 at Appendix D.

ensemble methods and item response theory.

Regarding *(5) score use and consequences*, the fairness analysis on the core corpus reveals no language, gender or task bias, suggesting models generalize well under strong rater agreement. Biases observed on weak agreement samples highlight areas for improvement. The absence of bias on the generalization corpus demonstrates that the models can handle different populations fairly. Moreover, the ECE study shows models are well-calibrated with fewer errors than human raters, reinforcing their reliability in high-stakes environments. Variability in human-rater agreement for correct and incorrect predictions emphasizes the need for cautious consideration. The varying trends in model performance based on rater agreement suggest some models might be better suited to specific contexts or types of essays.

8 Final Remarks

The main objective of our work was to address AES in the context of high-stakes language tests, specifically for FFL. This was motivated by the gap between existing AES evaluation standards and the high-stakes requirements of language certification. Our study focused on assessing the validity of AES models, emphasizing construct validity as well as raters' agreement and model generalizability. We systematically compared different AES modeling approaches using various evaluation metrics on two corpora (core and generalization) for a better study of generalizability. We analyzed AES models' predictions in terms of alignment with the gold standard, the indirect relation to linguistic features associated with language proficiency, fairness of treatment for test-takers, and the monitoring of failures in a high-stake environment.

Our results are supported by substantial resources. The core corpus is the largest calibrated corpus examined in French AES for certification and the generalization corpus benefits from numerous human raters, supporting reliable evaluation estimates. We found that metrics that overlook exact predicted levels (e.g. QWK and correlation) can give an appearance of high performance when models fail to predict exact levels. On the other hand, metrics focusing solely on exact level prediction (e.g. F1 and exact accuracy) fail to capture whether models understand the level scale. Additionally, model fairness must be evaluated beyond the entire corpus, including on cases where humans disagree.

We extended the standard AES ABV framework by operationalizing different abstract assessment areas with quantitative metrics, based on agreement, generalization, and construct validity. This bridges the gap between work on AES architectures and broader language assessment studies. Our study also advances the state of the art in French AES by proposing new models that outperform existing approaches. To the best of our knowledge, we also conducted the largest corpus study of linguistic features for FFL to date, providing valuable data for machine learning models and informing studies on teaching and assessing FFL.

Our study focused on a single language, training corpus, and language model. Nevertheless, the concatenation and baseline architectures align with those explored for English. In particular, the results for *SL* are encouraging and motivate further exploration with other languages. The limitation to one language model is due to the very limited number of models available for French, CamemBERT being the most widely used transformer encoder for French. Given the lack of comparative studies in French AES, we took a first step to fill this gap. Future research should compare architectures across diverse corpora for broader generalization.

In conclusion, our results are encouraging for applying AES models to high-stakes language certification tests in French. They also suggest that human raters consider elements beyond linguistic features captured in our AES models, stressing the complementarity between human and automated raters. FEI started using an automated scoring engine in conjunction with human raters. This is needed to closely and continuously monitor the model's performances and compensate for its shortcomings. For example, humans can detect outliers (such as off-topic answers), which the model cannot do. Another important aspect to monitor which might affect the model's performance is the constant renewal of prompts made by FEI.

For future work, we plan to study the characteristics of essays that all models fail to predict correctly. We also plan to investigate the generalizability to other tasks and languages as well as the impact of corpus size on model performance.

9 Acknowledgements

This research has been partially funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under the grant MIS/PGY F.4518.21

and T.0080.23, and also by a research convention with France Éducation International. Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.

References

AERA, APA, and NCME associations. 2014. *Standards for educational and psychological testing*. American Educational Research Association.

Y Attali. 2009. Evaluating automated scoring for operational use in consequential language assessment—the ets experience. In *annual meeting of the National Council on Measurement in Education, San Diego, CA*.

Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.

Randy Elliot Bennett and Isaac I Bejar. 1998. Validity and automad scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, 17(4):9–17.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Renske Bouwer, Anton Béguin, Ted Sanders, and Huub Van den Bergh. 2015. Effect of genre on the generalizability of writing scores. *Language Testing*, 32(1):83–100.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Stindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 1281—1288.

Brent Bridgeman, Catherine Trapani, and Yigal Attali. 2009. Considering fairness and validity in evaluating automated scoring. In *annual meeting of the National Council on Measurement in Education, San Diego, CA*.

Jill Burstein and Martin Chodorow. 1999. Automated essay scoring for nonnative english speakers. In *Computer mediated language assessment and evaluation in natural language processing*.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary Dee Harris. 1998. Automated scoring using a hybrid feature identification technique. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 206–210, Montreal, Quebec, Canada. Association for Computational Linguistics.

Carol A Chapelle, Mary K Enright, and Joan M Jamieson. 2008. *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Brian E Clauser, Michael T Kane, and David B Swanson. 2002. Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education*, 15(4):413–432.

Open Cambridge Learner Corpus. 2017. Distributed by lexical computing limited on behalf of cambridge university press and cambridge english language assessment.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Council of Europe Publishing.

Bastien De Clercq and Alex Housen. 2017. A cross-linguistic perspective on syntactic complexity in l2 development: Syntactic elaboration and diversity. *The Modern Language Journal*, 101(2):315–334.

Afrizal Doewes, Nughthoh Kurdhi, and Akrati Saxena. 2023. Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In *16th International Conference on Educational Data Mining, EDM 2023*, pages 103–113. International Educational Data Mining Society (IEDMS).

T. Eckes. 2009. *Quantitative Data Analysis for Language Assessment Volume I: Fundamental Techniques*, 1 edition. Routledge.

Fanny Forsberg Lundell and Christina Lindqvist. 2014. Vocabulary aspects of advanced l2 french: Do lexical formulaic sequences and lexical richness develop at the same rate? In *The Acquisition of French as a Second Language: New developmental perspectives*, pages 75–94. John Benjamins Publishing Company.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Jinyan Huang and Patrick B Whipple. 2023. Rater variability and reliability of constructed response questions in new york state high-stakes tests of english language arts and mathematics: implications for educational assessment policy. *Humanities and Social Sciences Communications*, 10(1):1–10.

Shi Huawei and Vahid Aryadoust. 2023. A systematic review of automated writing evaluation systems. *Education and Information Technologies*, 28(1):771–795.

Zhiwei Jiang, Tianyi Gao, Yafeng Yin, Meng Liu, Hua Yu, Zifeng Cheng, and Qing Gu. 2023. Improving domain generalization for prompt-aware essay scoring via disentangled representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12456–12470, Toronto, Canada. Association for Computational Linguistics.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. Tdnn: a two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1097.

Michael T Kane. 2013. Validating the interpretations and uses of test scores. *Journal of educational measurement*, 50(1):1–73.

Beata Beigman Klebanov and Nitin Madnani. 2021. Automated essay scoring. *Synthesis Lectures on Human Language Technologies*, 14(5):1–314.

Jason Sebastian Kusuma, Kevin Halim, Edgard Jonathan Putra Pranoto, Bayu Kanigoro, and Edy Irwansyah. 2022. Automated essay scoring using machine learning. In *2022 4th International Conference on Cybernetics and Intelligent System (ICORIS)*, pages 1–5. IEEE.

Benoit Lemaire and Philippe Dessus. 2001. A System to Assess the Semantic Content of Student Essays. *Journal of Educational Computing Research*, 24(3):305–320.

Guoxi Liang, Byung-Won On, Dongwon Jeong, Hyun-Chul Kim, and Gyu Sang Choi. 2018. Automated essay scoring: A siamese bidirectional lstm neural network architecture. *Symmetry*, 10(12):682.

Susan Lottridge, Chris Ormerod, and Amir Jafari. 2023. Psychometric considerations when using deep learning for automated scoring. *Advancing natural language processing in educational assessment*, pages 15–30.

Anastassia Loukina, Nitin Madnani, and Klaus Zechner. 2019. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pages 1–10.

Anastassia Loukina, Klaus Zechner, James Bruno, and Beata Beigman Klebanov. 2018. Using exemplar responses for training and evaluating automated speech scoring systems. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–12, New Orleans, Louisiana. Association for Computational Linguistics.

Nitin Madnani and Aoife Cahill. 2018. Automated scoring: Beyond natural language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1099–1109, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune bert for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162.

Daniel F McCaffrey, Jodi M Casabianca, Kathryn L Ricker-Pedley, René R Lawless, and Cathy Wendler. 2022. Best practices for constructed-response scoring. *ETS Research Report Series*, 2022(1):1–58.

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

Samuel Messick. 1990. Validity of test interpretation and use. Technical Report ETS-RR-90-11, ETS, Princeton, N.J.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Ricardo Muñoz Sánchez, David Alfter, Simon Dobnik, Maria Irena Szawerna, and Elena Volodina. 2024. Jingle BERT, jingle BERT, frozen all the way: Freezing layers to identify CEFR levels of second language learners using BERT. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 137–152, Rennes, France. LiU Electronic Press.

Diane Nicholls. 2003. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference*, volume 16, pages 572–581.

Nicholas Parslow. 2015. Automated Analysis of L2 French Writing: a preliminary study. Master's thesis. Publisher: Unpublished.

Pearson. 2019. Pte academic automated scoring. retrieved from: https://assets.ctfassets.net/yqwtwibiobs4/018RxttvPWsMkkGIQJ5Gg3/6f410437ceb2c6f2762fbcdfa8a28e8c/2021_PTEA_White_Paper_Institutions_Automated_Scoring_White_Paper-May-2018.pdf, accessed june 30th, 2024.

Chaitanya Ramineni, Catherine S. Trapani, David M. Williamson, Tim Davey, and Brent Bridgeman. 2012. Evaluation of the e-rater® scoring engine for the toefl® independent and integrated prompts. *ETS Research Report Series*, 2012(1):i–51.

Bojana Ranković, Sarah Smirnow, Martin Jaggi, and Martin J. Tomasik. 2020. Automated Essay Scoring in Foreign Language Students Based on Deep Contextualised Word Representations. In *LAK20-10th International Conference on Learning Analytics & Knowledge*. Issue: CONF.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. 2019. Language models and automated essay scoring. *arXiv preprint arXiv:1909.09482*.

Lawrence M. Rudner, Veronica Garcia, and Catherine Welch. 2006. An evaluation of IntelliMetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, 4(4).

Mark D. Shermis. 2015. Contrasting State-of-the-Art in the Machine Scoring of Short-Form Constructed Responses. *Educational Assessment*, 20(1):46–65.

Elana Shohamy, Smadar Donitsa-Schmidt, and Irit Ferman. 1996. Test impact revisited: Washback effect over time. *Language testing*, 13(3):298–317.

Steven E Stemler and Jessica Tsai. 2008. Best practices in interrater reliability three common approaches. *Best practices in quantitative methods*, pages 29–49.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.

Yi Tay, Minh Phan, Luu Anh Tuan, and Siu Cheung Hui. 2018. Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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

Masaki Uto, Itsuki Aomi, Emiko Tsutsumi, and Maomi Ueno. 2023. Integration of prediction scores from various automated essay scoring models using item response theory. *IEEE Transactions on Learning Technologies*.

Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088.

Salvatore Valenti, Francesca Neri, and Alessandro Cucchiarelli. 2003. An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, 2(1):319–330. Publisher: Informing Science Institute.

Alexander Von Eye and Eun Young Mun. 2014. *Analyzing rater agreement: Manifest variable methods*. Psychology Press.

Sara Cushing Weigle. 2013. English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, 18(1):85–99. Automated Assessment of Writing.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P. Yancey, and Thomas François. 2022. FABRA: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233, Marseille, France. European Language Resources Association.

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

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

David M Williamson, Xiaoming Xi, and F Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, 31(1):2–13.

Xiaoming Xi. 2008. What and how much evidence do we need? critical considerations in validating an automated scoring system. In C.A. Chapelle, Y.R. Chung, and J. Xu, editors, *Towards adaptive CALL: Natural language processing for diagnostic language assessment*, pages 102–114. Iowa State University Ames, IA.

Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. 2022. Automated essay scoring via pairwise contrastive regression. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2724–2733, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Duanli Yan and Brent Bridgeman. 2020. Validation of automated scoring systems. In *Handbook of Automated Scoring*, pages 297–318. Chapman and Hall/CRC.

Duanli Yan, André A Rupp, and Peter W Foltz. 2020. *Handbook of automated scoring: Theory into practice*. CRC Press.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569.

Tal Yarkoni, David Balota, and Melvin Yap. 2008. Moving beyond coltheart's n: A new measure of orthographic similarity. *Psychonomic bulletin & review*, 15(5):971–979.

Wajdi Zaghouani. 2002. AUTO-ÉVAL : vers un modèle d'évaluation automatique des textes. In *Actes du colloque des étudiants en sciences du*

langage, page 16, Montréal, Canada. Université du Québec à Montréal.

Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. 2015. Task-independent features for automated essay grading. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, pages 224–232.

A Hyperparameter selection

We explored the following hyperparameters: learn rate: 0.0005, 0.005, 1.00E-05, 2.00E-05, 3.00E-05, 4E-05, 5E-05; optimisation algorithm: adam, sgd; and the use of gradient clipping. The selected hyperparameter are presented in Table 4.

Model	**learn rate**	**optimisation**	**clip**
TR	3E-05	adam	no
CM	2E-05	adam	no
C2	2E-05	adam	no
SC	2E-05	adam	yes
IF	2E-05	adam	no
IV	2E-05	adam	no

Table 4: Selected hyperparameter for each model.

B Core Corpus Size

The corpus used in our study consists of 27,683 essays categorized across 6 CEFR levels (A1 to C2). Each essay is associated with 3 different tasks, with the total distribution shown in Table 5. The corpus is representative of a wide range of language of use, and covering essays from different prompts.

	Score	Task: descriptive	Task: narrative	Task: argumentative	#prompt	#language
A1	2883	306	969	921	993	82
A2	5678	593	1761	1772	2145	109
B1	5793	694	1850	1883	2060	126
B2	5881	721	2012	1987	1882	117
C1	5408	651	1727	1966	1715	104
C2	2040	486	657	716	667	54
Total	27683	856	8976	9245	9462	164

Table 5: Core corpus size

C Model performance and agreement between raters

Table 6 shows the average percentage agreement between raters, broken down by models performance (correct, mixed, wrong), models error (correct or wrong), and architecture.

group	**case**	**majority TR**	**majority SC**	**majority IV**	**majority SL**
all right	corr	78,17 %	78,17 %	78,17 %	78,17 %
mixed	corr	67,68 %	67,68 %	64,30 %	67,53 %
mixed	wrong	64,45 %	64,45 %	67,68 %	64,47 %
all wrong	wrong	62,81 %	62,81 %	62,81 %	62,81 %

Table 6: Average percentage agreement between raters for model correctness and error

D Details of the selected feature

The 48 features were selected as portrayed in Table 7. For each feature (column "Variable"), we report its family, its correlation with the CEFR levels in the core corpus, and the motivation behind its integration into the final set.

Family	Variable	Spearman	Motivation
Word length	LENwrdSYL_avg	0,6	Best Spearman, Pearson, mutual info gain across all levels
Sentence Length	LENsntWRD_q1	0,43	Best Spearman across all levels
	LENsntWRD_q3	*0,38*	Best mutual info gain + close to best Pearson
Lexical diversity	LEXdvrWSR_avg	0,64	Best Spearman, mutual info gain across all levels
	LEXdvrFLR_avg	*0,61*	Best C2 vs all FLC_avg
	LEXdvrNSM_avg	*0,44*	Best B1 vs all (-0.07)
	LEXdvrVSC_avg	*0,52*	Highly ranked across all families
Graded lexicons	LEXgrdFA1_20P	-0,61	Best Spearman across all levels
	LEXgrdFA2_20P	*-0,56*	The levels between A2 and C2 have been added to provide
	LEXgrdFB1_20P	*-0,51*	additional information for level A1.
	LEXgrdFB2_20P	*-0,5*	
	LEXgrdFC1_20P	*-0,39*	
	LEXgrdFC2_20P	*-0,29*	
	LEXgrdFMLA1_rsd	*0,52*	Best mutual info gain
Orthographic neighbors	LEXnghORT_80P	0,57	Best Spearman, mutual info gain across all levers
	LEXnghNUM_20P	*-0,55*	Best Pearson across all levels + A1 vs all, B2 vs all + A2 vs B1, B1 vs B2
	LEXnghPHO_avg	*0,52*	Best B1 vs all + B2 vs C1 (0,19)
Content Overlap	LEXcovLLST_90P	0,43	Best Spearman, Person across all levels
Lexical sophistication	LEXsopFK1_avg	0,47	Best Spearman across all levels
	LEXsopLLK3_max	*0,41*	Best A1 vs all + B2 vs all + different resource
Lexical Norms	LEXnrmAOA_max	0,4	Best Spearman and mutual info gain across all levels
	LEXnrmIMG_10P	*-0,36*	Best Pearson across all levels + different resource
	LEXnrmFAM_rsd	*0,2*	Best B1 vs all (0,1) , B2 vs C1, C1 vs C2 + different resource
Morphology features	LEXafxA_avg	0.48	Information diversity
Language development	SYNdevHGT_90P _max	0,52	_max Best Spearman, Pearson across all levels, 90P close
	SYNdevNPRSVPPART_avg	*0,31*	Best B1 vs all (0,08)
	SYNdevSIMA_90P	*0,41*	Best C1 vs C2 (0,08)
Dependency Relations	SYNdepMARK_90P	0,47	_max Best Spearman across all levels, 90P close
	SYNdepREPARANDUM_kurtosis	*-0,41*	Best Pearson across all levels
	SYNdepCOP_avg	*0,04*	Best C1 vs C2
	SYNdepACL_dolch	*0,37*	Best B1 vs all + _max B1 vs B2, B2 vs C1
POS Tag	SYNposADP_avg	0,31	Best Spearman across all levels
	SYNposSCONJ_skewness	0,31	Best Pearson across all levels
Clause features	SYNclsTYPCX_90P	0,46	Best Spearman across all levels
Morphology features	SYNmorVERBFORM_FIN_20P	-0,38	Best Spearman across all levels
	SYNmorNUMBER_PLUR_var	*0,33*	Best A1 vs A2 + interetsing information
	SYNmorVERBFORM_PART_90P	*0,27*	Best C1 vs C2 + information on time
Tense features	SYNtnsfINDP_avg	-0,42	Best Spearman, Pearson, MutInfo across all levels
	SYNtnsfPP_kurtosis	*0,22*	Best B1 vs all + information on time
	SYNtnsfTAP_skewness	*0,3*	Best C1 vs C2 (0,05) + information on time
Referential expressions	DISrefDN_avg	0,24	Best Spearman, Pearson across all levels
	DISrefPRSW_avg	*-0,19*	Best C1 vs all (0,11) + B1 vs B2 (-0,09)
Dialogue Variables	DISdiaPPEI2_skewness	*0,19*	Best Pearson, MutInfo across all levels + A2 vs B1, B1 vs B2
Text coherence	DIScohLSALADJ_90P	0,46	Best Pearson, MutInfo across all levels + close for Spearman across all levels + best B2 vs C1, C1 vs C2
Grammatical errors	ERRgrmSUM_avg	-0,73	Best Spearman, Pearson, MutInfo across all levels
	ERRlexSUM_avg	-0,32	Best Spearman , Pearson MutInfo across all levels
	ERRtypSUM_avg	-0,69	Best Spearman, Pearson, MutInfo across all levels
	ERRallSUM_avg	-0,79	Best Spearman, Pearson, MutInfo across all levels

Table 7: 48 selected features. For an explanation about them, see https://cental.uclouvain.be/fabra/docs_expert.html.

E Results of the direct comparison and linguistic correlation

Corpus	Model	Direct Comparison							Δ
		ρ	κ	QWK	MSE	EA	AA	$F1$	**Corr**
Core Corpus	TR	0.90	0.48	0.89	0.50	0.57	0.98	0.57	-0.02
	CM	0.90	0.48	0.89	0.49	0.57	0.98	0.57	-0.02
	C2	0.90	0.48	0.89	0.49	0.57	0.98	0.57	-0.02
	SC	0.90	0.44	0.88	0.55	0.54	0.97	0.54	-0.02
	IF	0.90	0.48	0.89	0.48	0.58	0.98	0.57	-0.02
	IV	0.90	0.48	0.89	0.49	0.57	0.98	0.57	-0.01
	SL	0.90	0.50	0.89	0.44	0.60	0.99	0.59	-0.03
	HL	0.90	0.47	0.88	0.48	0.57	0.98	0.56	-0.02
	Raters	0.91	0.70	0.91	0.40	0.75	0.98	0.75	0.01
Full agreement	TR	0.92	0.55	0.91	0.40	0.63	0.99	0.64	-0.01
	CM	0.92	0.56	0.91	0.39	0.64	0.99	0.64	-0.01
	C2	0.92	0.55	0.91	0.40	0.63	0.99	0.63	-0.01
	SC	0.92	0.51	0.90	0.45	0.60	0.98	0.60	-0.01
	IF	0.92	0.56	0.91	0.38	0.64	0.99	0.64	-0.01
	IV	0.92	0.55	0.91	0.39	0.63	0.99	0.64	-0.01
	SL	0.92	0.57	0.91	0.36	0.66	0.99	0.65	-0.02
	HL	0.91	0.56	0.91	0.38	0.64	0.99	0.63	-0.01
	Raters	1.00	1.00	1.00	0.00	1.00	1.00	1.00	-0.00
Close agreement	TR	0.89	0.41	0.87	0.59	0.52	0.97	0.51	-0.02
	CM	0.89	0.40	0.87	0.59	0.51	0.97	0.50	-0.02
	C2	0.89	0.42	0.87	0.56	0.53	0.97	0.51	-0.02
	SC	0.89	0.37	0.86	0.65	0.48	0.96	0.47	-0.02
	IF	0.89	0.41	0.87	0.57	0.52	0.97	0.50	-0.02
	IV	0.89	0.41	0.87	0.57	0.52	0.97	0.50	-0.02
	HL	0.89	0.45	0.88	0.50	0.55	0.98	0.53	-0.03
	SL	0.88	0.40	0.86	0.56	0.52	0.98	0.49	-0.02
	Raters	0.87	0.40	0.86	0.61	0.51	0.99	0.50	0.01
Weak agreement	TR	0.84	0.23	0.82	0.87	0.36	0.92	0.36	-0.05
	CM	0.84	0.21	0.82	0.86	0.34	0.94	0.35	-0.05
	C2	0.83	0.24	0.82	0.86	0.37	0.93	0.36	-0.05
	SC	0.83	0.24	0.82	0.91	0.37	0.91	0.37	-0.06
	IF	0.83	0.23	0.81	0.85	0.36	0.93	0.34	-0.05
	IV	0.83	0.24	0.82	0.85	0.37	0.93	0.35	-0.05
	SL	0.83	0.21	0.81	0.84	0.35	0.94	0.33	-0.06
	HL	0.82	0.18	0.80	0.88	0.32	0.94	0.30	-0.06
	Raters	0.67	0.33	0.66	1.88	0.45	0.57	0.44	0.06

Table 8: Results of the direct comparison and linguistic correlation analysis in the core corpus, including its splits by human raters agreement.

F Correlations Between Candidates' Errors and Model Residuals

CEFR	#docs	Lexical	Morphological	Syntactic
A1	28	0.35**	0.15	0.14
A2	61	0.25**	0.14	0.14
B1	66	0.14	0.08	0.08
B2	68	0.08	0.04	0.05
C1	61	0.05	0.02	0.03
C2	24	0.03**	0.01	0.02*

(a) Proportion of errors per essay

CEFR	#docs	Lexical	Morphological	Syntactic
A1	28	−0.53**	−0.08	−0.28
A2	61	−0.45**	−0.19	−0.20
B1	66	−0.03	−0.03	−0.10
B2	68	−0.07	−0.03	−0.05
C1	61	0.09	0.04	0.20
C2	24	0.62**	0.08	0.46*

(b) Correlations between presence of candidates' errors in the texts

Table 9: Error annotation in the *soft-labeling* model's residuals, by CEFR level and error type. Correlations with p-value ≤ 0.05 are marked with * and ≤ 0.01 are marked with **.

G Examples of essays that challenge the model

Level		Essay	
Model	**Humans**	**Original**	**English Translated**
A1	B1	bonjour nicolas comme va tu . de mon cote sa ne va pas deus voles sont entre chez moi et il ont vole mont sanc il ya tout mes papiers j'ai porte pleinte a la gandameri menteman j'ai plus de passeport ni de cate d'idandite bonne journne a toi et je te tien ou courent si la polise	hello nicolas, how are you. as for me things are not going well two robbers came into my house and stole my backpack there is all my papers I filed a complaint with the police I do not have a passport or ID card have a good day and I'll let you know if the police.
A1	B1/B2	bonjours chèrès intèrnautès jèspèrs què vous allèr bièn. jai ètè invitè lè wèènkènd passè lors du mariagè organisè par la famillè . jè suis hypèr èxcistè je vous fairè part dè cèttè cèromoniè car cè qui ma intèrrèsè dunè part il yavait lès joliè dècorations ,idèniablè , dè la boissons dè lambiancès jusqua laubè la mariè ètait bèllè oh lalal ainsi què celui du mariè.ènsuitè ,cè mariagè ètait spèstaculairè parmis tans dè cèrèmoniè què jai ètait invitè.pour cèla je vous invitè a assitèr lès cèrèmoniès qui sèront important pou lè bièn ètrè dè votrè èpanouissèmènt jè vouys rèmèrciè a bièntot.	hello dear internet users, I hope you are well. I was invited last weekend to the wedding organized by the family. I'm super excited, I'm sharing this ceremony with you because what interested me on the one hand there were the pretty decorations, idèniable, from the party drinks until dawn the bride was beautiful oh lalal as well as that of the bridegroom. Then, this wedding was spectacular among so many ceremonies that I was invited. For this I invite you to attend the ceremonies which will be important for your well-being.I thank you see you soon.

Table 10: Examples of challenging essays

H Fine-grained inspection of the results

	A1	A2	B1	B2	C1	C2
TR	0.66	0.62	0.57	0.51	0.55	0.54
CM	0.62	0.62	0.57	0.53	0.54	0.54
C2	0.65	0.63	0.55	0.54	0.56	0.52
SC	0.62	0.61	0.54	0.49	0.48	0.48
IF	0.62	0.64	0.56	0.54	0.56	0.48
IV	0.62	0.63	0.56	0.52	0.57	0.53
SL	0.65	0.63	0.58	0.58	0.62	0.47
HL	0.58	0.60	0.57	0.55	0.62	0.44
Raters	0.76	0.76	0.74	0.75	0.76	0.74
SL-GC	0.57	0.59	0.71	0.74	0.71	0.68

Table 11: F1 by CEFR level on the Core corpus and the SL model on the Generalization corpus (GC).