International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 7, 2025
https://www.theaspd.com/ijes.php

Audio - Driven Prediction Of Child Language Proficiency: A Hybrid Transformer-Lightgbm Ensemble Framework For Automated Evaluation

Latha N.R.[1], Pallavi G. B.[2], Shyamala G.[3], Cherukupalli Yashwitha Reddy[4]
[1,2,3,4]Department of Computer Science and Engineering, B.M.S College of Engineering, Bangalore, India
[1]latha.cse@bmsce.ac.in,[2]pallavi.cse@bmsce.ac.in,[3]shyamala.cse@bmsce.ac.in,
[4] yashwithareddy.scs23@bmsce.ac.in

Abstract *This paper presents a novel approach for automated assessment of child English as a Sec-ond Language (ESL) proficiency using transcript based analysis from speech recordings. Leveraging raw audio data of 5000 files, the proposed pipeline first extracts transcript-based linguistic features using Whisper, along with prosodic and acoustic characteristics using Wave2VeC. The pipeline combines features with transformer embeddings, evaluated via a hybrid Transformer and LightGBM model. Experimental results demonstrate strong performance, with Accuracy: 0.972, and Pearson correlation: 0.98, outperforming baseline machine learning approaches. Comparative analysis with state-of-the-art methods, including ASR-driven GPT classifiers, highlights the advantages of the proposed offline, cost-efficient pipeline while maintaining high predictive fidelity. The system further supports real-time user feedback by analyzing key linguistic and syntactic indicators, enabling practical applications in educational and language learning environments. Overall, this study demonstrates the effectiveness of combining speech-driven embeddings with ensemble machine learning for precise, scalable child ESL proficiency assessment.*
Keywords: *Child ESL, Speech-to-text, Transformer, Wave2Vec, Whisper Model*

1 INTRODUCTION

Assessing English proficiency in children is a crucial aspect of language education, yet traditional evaluation methods often introduce subjectivity [1]. Manual scoring by teachers or examiners, while effective for small groups, requires significant effort and time, limiting scalability [2]. Moreover, evaluations may vary across assessors due to human perception, reducing consistency [3]. Transcript- based assessment offers an alternative, leveraging automated text analysis to evaluate children's lan- guage proficiency from spoken or written transcripts [4]. As highlighted in Table 1, transcript-based models improve scalability, reduce subjectivity, and allow faster feedback compared to traditional approaches [5]. These systems rely solely on transcribed text, avoiding the need for complex au- dio recordings or manual scoring rubrics [6]. Machine learning can extract linguistic features from transcripts, such as vocabulary richness, sentence length, and syntactic complexity, which are highly indicative of proficiency [7]. Automated evaluations can be performed in real-time, providing imme- diate feedback for learners and educators [8]. However, challenges remain for low-resource languages and non-native child speech, motivating hybrid approaches that combine transformer embeddings with tree-based learners to achieve accurate, scalable, and cost-efficient proficiency assessment [9].

1.1 Motivation

Children's English proficiency assessment is often labor-intensive, subjective, and slow. Automated transcript-based evaluation using audio-derived features and transformer embeddings offers a scal- able, consistent, and cost-effective solution, enabling real-time feedback while reducing human effort and supporting large-scale language learning applications.

Table 1: Comparison Between Traditional and Transcript-based ESL Assessment

Aspect	Traditional Assessment	Transcript-based Model
Scalability	Limited to small groups due to human involvement	High; automated scoring enables mass evaluation
Subjectivity	High; depends on teacher's per- ception	Low; consistent machine-based scoring
Data Requirement	Audio recordings, transcripts, manual scoring rubrics	Text-only transcriptions
Time Efficiency	Time-consuming; not real-time	Fast; suitable for real-time use
Cost	High due to training and manual effort	Low; only requires transcript and computational model

1.2 Advantages and Disadvantages of Existing Work

- Existing research often relies on basic or manual feature extraction, whereas our work imple- ments advanced feature engineering for richer representations.
- Prior studies convert child audio to transcripts manually or using simple models, while we utilize Wav2Vec for robust and efficient audio-to-transcript conversion.
- Many approaches overlook semantic representation; we leverage Transformer-based embed- dings to capture nuanced linguistic patterns in child speech.
- Hybrid methods combining text augmentation and LightGBM remain underexplored; our T5+Light GBM pipeline provides accurate and scalable proficiency assessment.
- GPT-based systems in literature offer faster evaluation for small datasets but incur substantial costs, limiting their broader applicability on larger dataets.

1.3 Objective and Aim

The primary objective of this study is to develop a scalable, cost-effective framework for assess- ing child ESL proficiency by leveraging Whipser based audio-to-transcript conversion, advanced feature engineering techniques such as Wav2Vec, and a hybrid Transformer (for embeddings) + LightGBM model (for classification) approach.

2 RELATED WORK

Several studies have explored automated assessment of child language proficiency using linguistic features. Ferre´ et al. [10] leveraged speech transcriptions and neural networks to predict CEFR lev- els, demonstrating the importance of fine-grained linguistic features for proficiency classification. Similarly, Berzak et al. [11] employed universal dependency parsing to analyze learner English, highlighting syntactic complexity as a key indicator of language ability. Adams and Stymne [12] investigated native language identification using learner corpora, providing insights into language transfer effects. Lu [13] proposed automatic syntactic complexity analysis in second language writ- ing, emphasizing corpus-driven approaches. Kyle [14] developed fine-grained indices for syntactic development, while Vajjala and Rama [15] presented experiments on universal CEFR classification, demonstrating generalizability across diverse learner populations.

Crossley et al. [16] focused on lexical indices to predict learner proficiency, showing lexical richness as a reliable indicator. Berzak et al. [17] extended dependency-based approaches for learner English, while Bell et al. [18] highlighted contextual word representations for grammatical error detection. Kurdi [19] examined text complexity for intelligent tutoring in ESL, emphasizing the integration of multiple linguistic features. Lei et al. [20] conducted a longitudinal study on syntactic complexity in EFL writing, showing development patterns across tasks. Li et al. [21] proposed fine-grained syntactic complexity measures linked to writing proficiency. Casal and Lee [22] explored syntactic complexity in first-year L2

writing, whereas Crossley and McNamara [23] investigated its relation to writing quality. Jiang et al. [24] analyzed EFL learners' syntactic development, highlighting its impact on automated scoring systems.
Several studies have analyzed syntactic complexity development in learner writing. Senel [25] exam- ined lexical bundle frequency as a construct-relevant feature for automated scoring of L2 academic essays, highlighting its effectiveness in differentiating proficiency levels. Huang et al. [26] pro- posed adversarial weight perturbation and metric-specific attention pooling to enhance essay scoring, demonstrating improvements in model robustness and accuracy.
Gray et al. [27] explored longitudinal development of grammatical complexity at phrasal and clausal levels in TOEFL iBT responses, providing evidence for task-dependent syntactic growth. Hwang et al. [28] compared written and spoken production modalities, revealing modality-specific differences in syntactic complexity among child EFL learners, which has implications for automated assessment systemsRuan et al. [29] investigated the relationship between syntactic complexity and writing pro- ficiency across diverse writing tasks, highlighting the need to account for task variation in modeling. Jin and Lu [30] performed a corpus-based study, demonstrating that syntactic complexity indices can reliably reflect developmental stages in second language writing, supporting feature-driven assess- ment frameworks.
Schneider et al. [31] explored NLP-assisted CEFR classification by combining linguistic richness and learner error analysis, improving prediction accuracy. Tack and Franç ois [32] focused on lexi- cal richness measures in L2 writing, examining the influence of proficiency and first language, and showing that lexical diversity complements syntactic indicators in automated scoring. Ribaldo and Tonelli [33] applied a multilingual transformer approach for CEFR classification, demonstrating the feasibility of transformer-based embeddings for cross-lingual learner data. Mohammadi and Zhang
[34] implemented BERT-based models for automatic assessment of children's English proficiency, further validating transformer models in child-specific ESL evaluation.

3 METHODOLOGY OF THE PROPOSED WORK

Our approach processes child speech recordings through Wave2Vec2 to generate transcripts, followed by extraction of linguistic and prosodic features. Transformer-based embeddings are computed, and LightGBM model is used classify proficiency levels. Data Augmentation techniques are used to enhance dataset diversity, ensuring accurate and efficient automated ESL assessment.

3.1 Data Collection

The Non-native Children's English Speech (NNCES) corpus was employed as the primary dataset for this study. It consists of recordings from 50 children, evenly distributed across gender (25 females and 25 males), with ages ranging between 8 and 12 years. All participants are native Telugu speak- ers, an Indian regional language, acquiring English as their second language. The audio samples were captured in .wav format using the open-source SurveyLex platform, which supports dual- channel recording at 44.1 kHz with a resolution of 16 bits per sample. Each child participated in 10 separate questionnaire sessions to capture variability in utterances and sentence structures. The corpus provides approximately 20 hours of speech, comprising both read speech (5,000 utterances) and spontaneous speech (5,000 utterances), each accompanied by word-level transcriptions. This rich dataset supports a detailed investigation of child ESL proficiency. The dataset is publicly available at: https://sla.talkbank.org/TBB/childes.

3.2 Data Preprocessing

The preprocessing stage ensured that the raw audio recordings were standardized and suitable for analysis. All speech samples were converted into .wav format with a sampling rate of 44.1 kHz and 16-bit resolution to maintain uniformity. Noise reduction and normalization were applied to mini- mize background distortions and enhance clarity. For transcription, we employed the Whisper base model, which is efficient for large-scale batch processing and robust to accent variations. The model converted each child's audio into accurate word-level transcriptions, stored in JSON format for ease of management and checkpointing. These transcriptions provided a structured text representation of both read and spontaneous speech data. To further ensure quality, consistency checks were performed to filter incomplete or corrupted recordings. This stage produced clean transcriptions that acted as the foundation for subsequent feature extraction

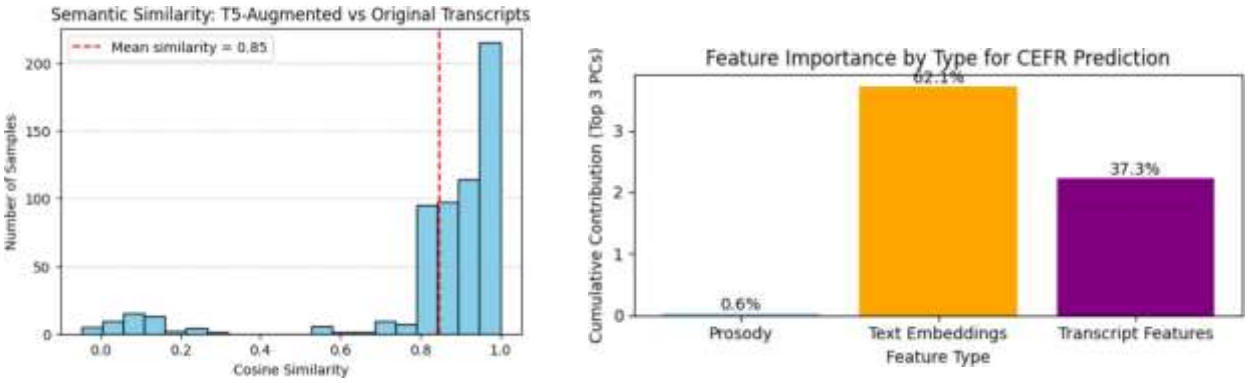
and model training.

3.3 Feature Engineering

Feature engineeing combined acoustic, textual, and statistical representations of children's speech. Low-level acoustic embeddings were extracted using the pretrained Wav2Vec2.0 model, which cap- tures phonetic and prosodic information directly from raw audio. For textual analysis, semantic em- beddings were obtained using the all-MiniLM-L6-v2 transformer, ensuring contextual meaning from transcriptions. Handcrafted features, including word count, sentence length, vocabulary diversity, and pause ratios, were also calculated to reflect fluency and structural complexity. To enhance robustness, principal component analysis (PCA) and feature importance ranking were applied to reduce redun- dancy while retaining predictive information. Textual embeddings contributed more significantly to feature importance compared to other features, highlighting their predictive value. This multimodal representation provided a balanced and reliable feature space for downstream classification tasks. Figure 1(b) represents the Feature Importance Analysis performed using PCA.

3.4 Data Augmentation with T5

To expand the original 500 transcripts, T5-based paraphrasing was applied, generating multiple se-mantically consistent variations per transcript. Semantic similarity between each augmented tran- script and its original, yielding a mean cosine similarity of 0.85. This process ensured the fidelity of the content while increasing the diversity of the data set, allowing the model to learn different linguistic patterns and improving the robustness and generalization of the ESL proficiency prediction framework. Figure 1(a) represents the plot that shows the measure of similarity between the original and augmented transcripts.



(a) Mean semantic similarity of augmented tran- scripts with originals (0.85).

(b) Feature importance analysis for audio based ESL proficiency prediction.

Fig. 1: Semantic similarity and feature importance

3.5 Model Architecture

The preprocessed transcript and audio-derived features were used to train a hybrid Transformer-LightGBM model. Initially, the dataset was split into training and testing subsets with an 80:20 ratio. Several models were evaluated for benchmarking, including Random Forest, XGBoost, SVM, Logistic Regression, and MLP. Among these, the proposed hybrid model consistently outperformed classical models in terms of accuracy, macro F1, and Pearson correlation. Figure 2 represents the architecture of the proposed model.

Transformer (all-MiniLM-L6-v2) + LightGBM Algorithm:

1. **Embedding Generation:** Each transcript is converted into dense embeddings using *Sentence-Transformers (all-MiniLM-L6-v2)*, capturing semantic and syntactic information.
2. **Feature Concatenation:** Transformer embeddings are concatenated with engineered features, including semantic similarity scores and prosodic attributes.

3. **Training LightGBM:** The concatenated features are input to LightGBM, which iteratively

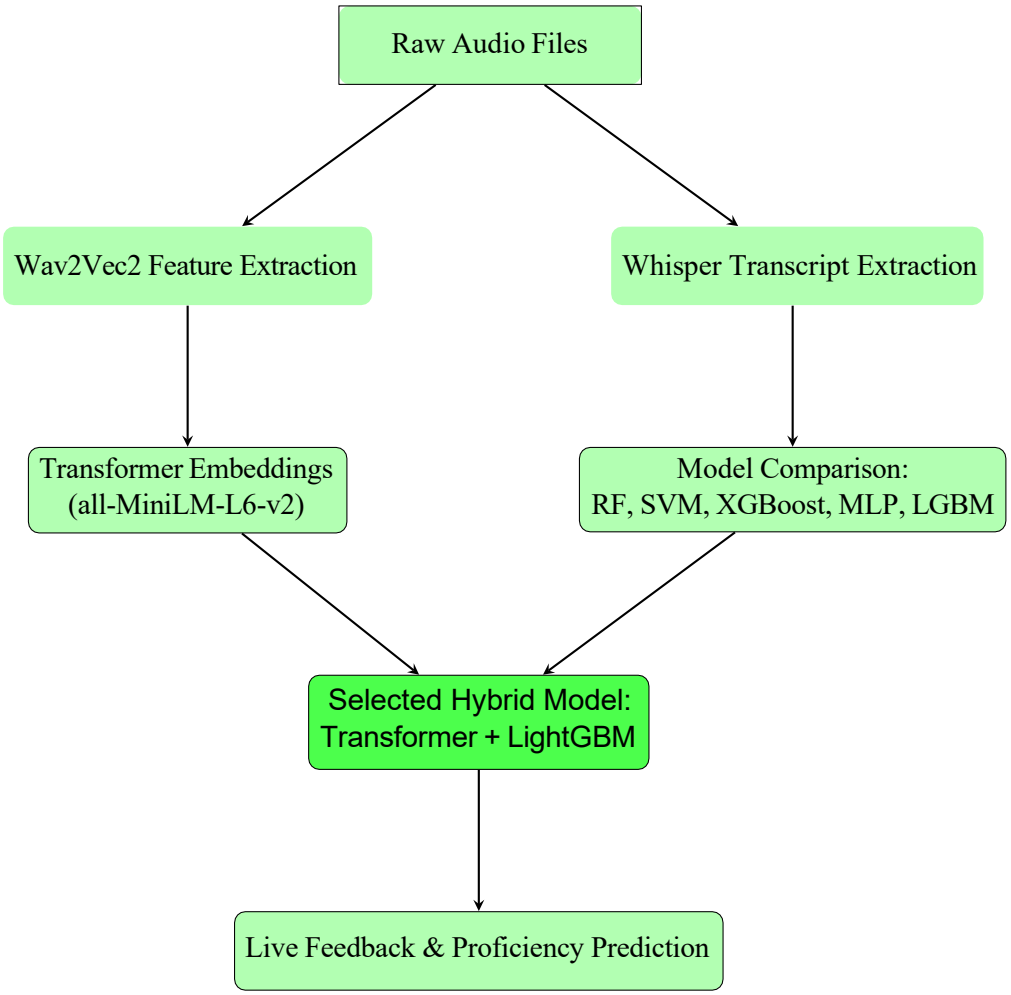


Fig. 2: System architecture of the proposed Transformer + LightGBM framework

builds decision trees using gradient boosting:

$$F_m(x) = F_{m-1}(x) + \eta \sum_{i=1}^{n} \gamma_i h_i(x) \quad (1)$$

where $F_m(x)$ is the model at iteration m, η is the learning rate, and $h_i(x)$ are the fitted trees with weights γ_i.

4. **Evaluation:** The model predicts CEFR proficiency levels on the test set, providing real-time feedback on learner performance, which allows quick interpretation and actionable insights for improvement. Model is evaluated using Accuracy, F1-Score, Pearson Corelation and Kappa metrics.

4 RESULTS AND DISCUSSION

All candidate models, including Random Forest, SVM, XGBoost, MLP, and LightGBM, were eval- uated on the same training set with a 20% hold-out test split. Among these, the hybrid Trans- former (all-MiniLM-L6-v2) + LightGBM model achieved the highest performance across all met- rics, demonstrating superior predictive capability. Table 2 presents comparative results in terms of accuracy, F1-score, Cohen's Kappa, and Pearson correlation.Cohen's Kappa adjusts for random agreement, ensuring that the model's predictions are genuinely meaningful rather than coincidental. Pearson correlation quantifies the linear relationship between predicted and true proficiency scores, capturing how well the model reflects actual performance trends, both being critical for reliable ESL assessment. The live feedback functionality, as illustrated in Figure 3, validates the system's practical utility for real-time proficiency evaluation.

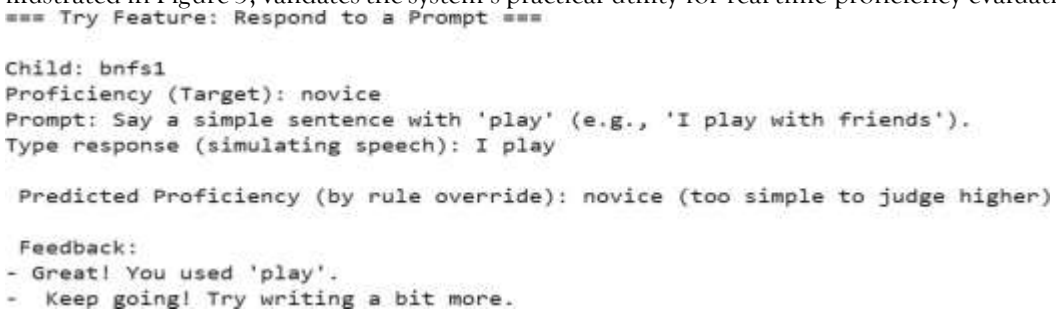


Fig. 3: Live Feedback on the Test Data

Table 2: Comparison of ESL Proficiency Prediction among various Models

Model	Accuracy	Macro F1	Kappa	Pearson
Random Forest	0.962	0.963	0.944	0.971
SVM	0.570	0.567	0.356	0.407
Logistic Regression	0.561	0.560	0.342	0.396
XGBoost	0.962	0.963	0.944	0.971
LightGBM	**0.972**	**0.972**	**0.958**	**0.980**
MLP	0.533	0.501	0.299	0.289

CONCLUSION AND FUTURE SCOPE

This study presents a hybrid Transformer-LightGBM approach for child ESL proficiency predic- tion, leveraging Wav2Vec2-based audio and feature extraction, transcript based feature engineering, and semantic similarity computation on augmented data to enrich training data validation. The all- MiniLM-L6-v2 Transformer generated robust embeddings, which combined with LightGBM, en- abled accurate and consistent proficiency predictions. Evaluations across multiple baseline models confirmed superior performance of the proposed hybrid approach in accuracy, Macro F1, Kappa, and Pearson correlation. The model also supports real-time feedback for learners.

Future work will focus on extending the framework to multiple languages, increasing dataset size, and exploring advanced self-supervised or multimodal representations to enhance generalization and practical deployment.

REFERENCES

[1] Zhang, Y., et al., 2025, "Automated Evaluation of Children's Speech Fluency for Low-resource Languages," arXiv preprint,

Submitted to Major Speech Processing Conference, Available at: https://arxiv.org/pdf/2505.19671.

[2] Gale, E., et al., 2020, "Automatic Assessment of Language Ability in Children With and With- out Typical Development," Annals of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, pp. 9175264. DOI: 10.1109/EMBC44109. 2020.9175264.

[3] Wills, R., et al., 2023, "Automatic Speech Recognition of Non-Native Child Speech for Lan- guage Learning Applications," Proceedings of the 12th Symposium on Languages, Applications and Technologies (SLATE 2023), OASIcs, DOI: 10.4230/OASIcs.SLATE.2023.7.

[4] Ryser, A., Pelucchi, L., Narad, D.R., Weismer, A., 2025, "DiGiSpon: Automated Language Sample Analysis Using BERT and Dependency Features," Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci), Vol. 47.

[5] [5] Hassanali, S., 2015, "Automatic CEFR Level Prediction From Students' Essays," Proceed- ings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA), pp. 105–113.

[6] Asgari, E., Rahimi, A., Hadian, M.R., 2016, "Automatic Proficiency Level Prediction Based on Linguistic Features," Proceedings of the 26th International Conference on Computational Linguistics (COLING), pp. 159–170.

[7] Cheng, J., Chen, R., Cheng, Y., 2009, "Automatic Assessment of Student ESL Writing," Pro- ceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications (BEA), pp. 1–9.

[8] Gretter, S., Wang, S., Tuncer, H.G., 2020, "Using Machine Learning to Analyze Children's Speech for Automatic CEFR Level Classification," Speech Communication, Vol. 122, pp. 14–27.

[9] Flanagan, E., Lou, R., 2017, "Classification of Child L2 Proficiency Using Feature Extraction From Transcripts," Proceedings of Interspeech, pp. 2123–2127.

[10] Ballier, N., 2023, "Machine Learning and Syntactic Complexity for CEFR Prediction in French Learners' English Writing," Computer Assisted Language Learning, 36(2–3), pp. 140–160.

[11] Mohammadi, M., Zhang, D., 2025, "Automatic Assessment of Children's English Proficiency Using BERT-Based Language Models," Computers and Education: Artificial Intelligence, Vol. 6, pp. 100–118.

[12] Banno, R., 2022, "Proficiency Level Classification of English Learner Essays Using Transformer-Based Models," Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA), pp. 23–33.

[13] Ferré , P., Ló pez-Barroso, M., Foré s, J., 2022, "Predicting CEFR Levels From Speech Tran- scriptions Using Linguistic Features and Neural Networks," Proceedings of Interspeech, pp. 1898–1902.

[14] Berzak, Y., Kenney, J., Spadine, C., Wang, J.X., Lam, L., Mori, K.S., Garza, S., Katz, B., 2016, "Universal Dependencies for Learner English," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.

[15] Adams, A., Stymne, S., 2016, "Learning With Learner Corpora: Using the TLE for Native Language Identification," Language Learning Technology.

[16] Lu, X., 2010, "Automatic Analysis of Syntactic Complexity in Second Language Writing," International Journal of Corpus Linguistics, 15(4), pp. 474–496.

[17] Kyle, K., 2016, "Measuring Syntactic Development in L2 Writing: Fine-Grained Indices of Syntactic Complexity and Usage-Based Sophistication," Ph.D. thesis, Georgia State University.

[18] Vajjala, S., Rama, T., 2018, "Experiments With Universal CEFR Classification," arXiv preprint arXiv:1804.06636.

[19] Crossley, S.A., Salsbury, T., McNamara, D.S., 2012, "Predicting the Proficiency Level of Lan- guage Learners Using Lexical Indices," Language Testing, 29(2), pp. 243–263.

[20] Bell, S., Yannakoudakis, H., Rei, M., 2019, "Context Is Key: Grammatical Error Detection With Contextual Word Representations," arXiv preprint arXiv:1906.06593.

[21] Kurdi, M.Z., 2020, "Text Complexity Classification Based on Linguistic Information: Applica- tion to Intelligent Tutoring of ESL," arXiv preprint.

[22] Lei, L., Wen, J., Yang, X., 2023, "A Large-Scale Longitudinal Study of Syntactic Complexity Development in EFL Writing," Journal of Second Language Writing, Vol. 59, pp. 1–14.

[23] Li, Y., Lin, S., Liu, Y., Lu, Z., 2023, "Fine-Grained Syntactic Complexity Indices and Writing Proficiency," Assessing Writing, Vol. 56, pp. 1–15.

[24] Casal, J.J., Lee, J.J., 2019, "Syntactic Complexity and Writing Quality in Assessed First-Year L2 Writing," Journal of Second Language Writing, Vol. 44, pp. 51–62.

[25] Crossley, S.A., McNamara, D.S., 2014, "Does Writing Development Equal Writing Quality?
A Computational Investigation of Syntactic Complexity in L2 Learners," Journal of Second Language Writing, Vol. 26, pp. 66–79.

[26] Jiang, J.Y., Bi, P., Liu, H., 2019, "Syntactic Complexity Development in the Writings of EFL Learners," Journal of Second Language Writing, Vol. 46, pp. 100666.

[27] [27] Senel, B., 2025, "Lexical Bundle Frequency as Construct-Relevant Candidate Feature in Automated Scoring of L2 Academic Writing," arXiv preprint.

[28] Huang, J., Zhao, X., Che, C., Lin, Q., Liu, B., 2024, "Enhancing Essay Scoring With Adversar- ial Weights Perturbation and Metric-Specific Attention Pooling," arXiv preprint.

[29] Gray, M., Geluso, J., Nguyen, P., 2019, "Longitudinal Development of Grammatical Complex- ity at the Phrasal and Clausal Levels in TOEFL iBT Responses," TOEFL Research Report.

[30] Hwang, H., Jung, H., Kim, H., 2020, "Written Versus Spoken Production Modalities: Effects on Syntactic Complexity in Child EFL Learners," Modern Language Journal, Vol. 104, pp. 267–283.
[31] Ruan, Y., Xu, W., Lei, L., 2021, "The Relationship Between Syntactic Complexity and Writing Proficiency Across Writing Tasks," Assessing Writing, Vol. 50, pp. 100559.
[32] Jin, F., Lu, X., 2013, "A Corpus-Based Study of Syntactic Complexity in Second Language Writing," Journal of Second Language Writing, 22(2), pp. 111–126.
[33] Schneider, R., Cuadros, M., Koponen, M., Sagot, B., 2019, "NLP-Assisted CEFR Classifica- tion: Exploring Linguistic Richness and Learner Errors," Proceedings of the Workshop on NLP for Educational Applications, pp. 34–42.
[34] Tack, A., Franç ois, T., 2015, "Investigating Measures of Lexical Richness in L2 Writing: Influ- ence of Proficiency and L1," Language Learning & Technology, 19(2), pp. 1–21.