Asia Pacific Journal of Corpus Research
Vol. 5, No. 1, 23-36

http://www.icr.or.kr/apjcr
e-ISSN 2733-8096
https://doi.org/10.22925/apjcr.2024.5.1.23

Corpus of Eye Movements in L3 Spanish Reading: A Prediction Model

Hui-Chuan Lu
(National Cheng Kung University)
Li-Chi Kao
(National Cheng Kung University)
Zong-Han Li
(National Cheng Kung University)
Wen-Hsiang Lu
(National Cheng Kung University)
An-Chung Cheng
(University of Toledo)

Lu, H. C., Kao, L. C., Li, Z. H., Lu, W. H., & Cheng, A. C. (2024). Corpus of eye movements in L3 Spanish reading: *A prediction model. Asia Pacific Journal of Corpus Research, 5*(1), 23-36.

This research centers on the Taiwan Eye-Movement Corpus of Spanish (TECS), a specially created corpus comprising eye-tracking data from Chinese-speaking learners of Spanish as a third language in Taiwan. Its primary purpose is to explore the broad utility of TECS in understanding language learning processes, particularly the initial stages of language learning. Constructing this corpus involves gathering data on eye-tracking, reading comprehension, and language proficiency to develop a machine-learning model that predicts learner behaviors, and subsequently undergoes a predictability test for validation. The focus is on examining attention in input processing and their relationship to language learning outcomes. The TECS eye-tracking data consists of indicators derived from eye movement recordings while reading Spanish sentences with temporal references. These indicators are obtained from eye movement experiments focusing on tense verbal inflections and temporal adverbs. Chinese expresses tense using aspect markers, lexical references, and contextual cues, differing significantly from inflectional languages like Spanish. Chinese-speaking learners of Spanish face particular challenges in learning verbal morphology and tenses. The data from eye movement experiments were structured into feature vectors, with learner behaviors serving as class labels. After categorizing the collected data, we used two types of machine learning methods for classification and regression: Random Forests and the k-nearest neighbors algorithm (KNN). By leveraging these algorithms, we predicted learner behaviors and conducted performance evaluations to enhance our understanding of the nexus between learner behaviors and language learning process. Future research may further enrich TECS by gathering data from subsequent eye-movement experiments, specifically targeting various Spanish tenses and temporal lexical references during text reading. These endeavors promise to broaden and refine the corpus, advancing our understanding of language processing.

Keywords: Corpus Development, Machine Learning, Input Processing, Second Language Acquisition, Eye-Tracking

1. Introduction

Research on the initial stages of the second language acquisition (SLA) processes, utilizing eye-tracking technology, has been gaining traction in Taiwan. For example, Liu et.al (2019) investigated the reading behavior of Taiwanese college English learners using eBooks to study the advantages of digital content in English learning. Similarly, Guan (2020) employed eye-tracking technology to examine how phonological processing influences Chinese speakers learning Japanese as their second language. Eye-tracking technology provides real-time insights into cognitive processes during language comprehension, making it an invaluable tool for SLA. Since 2015, we have also centered on conducting eye movement experiments to investigate various linguistic aspects, such as collocations, clitic pronouns, relative clauses, subjunctive mood, and the mapping of tenses and temporal adverbs in the acquisition of Spanish as a third language (L3) among learners in Taiwan. By detecting and tracking eye movements, we aim to unravel the input processing patterns and strategies employed by Spanish learners during reading activities. Through meticulous analysis of collected eye movement data, we have examined factors such as language proficiency level, vocabulary frequency, parts of speech, and position of the targeted linguistic elements at the sentence level. Previous research has demonstrated that eye-tracking can reveal patterns of attention and processing in language learners. Rayner's (1998) work on eye movements during reading laid the groundwork for understanding how native speakers process text, emphasizing the importance of fixation duration and saccades in comprehension. Dussias (2010) extended these findings to bilinguals, showing that eye-tracking can differentiate between proficient and less proficient language users based on their eye movement patterns. Building on these foundational studies, our research utilizes eye-tracking to investigate specific linguistic features in the acquisition of Spanish as a third language (L3) among Taiwanese learners. Additionally, studies such as those by Godfroid and Winke (2015) have demonstrated that eye-tracking can reveal how learners process grammar and vocabulary in real-time, offering a more nuanced understanding of SLA processes. In 2022, we established the Taiwan Eye-Movement Corpus of Spanish (TECS), a corpus of eye-tracking data, marking the first of its kind in Taiwan (Figure 1). This corpus stands as a testament to our commitment to sharing the valuable resources we have accumulated, thereby and it fosters collaborative research across diverse fields. We aim to deepen our understanding of the L3 Spanish acquisition of various linguistic features, with the goal of developing a comprehensive database for Spanish language acquisition. Our current research builds on the established theoretical framework by examining how eye-tracking data can enhance our understanding of L3 acquisition. Specifically, we draw on Gass and Mackey's (2007) interactionist approach, which posits that language learning is facilitated through interaction and attention to input. Eye-tracking data provides empirical evidence of where learners focus their attention, supporting theories of input processing (VanPatten, 2007). By doing so, we endeavor to furnish a theoretical foundation that can support teaching practices, inform the design of instructional materials, foster the integration of teaching and research, and ultimately advance knowledge in the field.

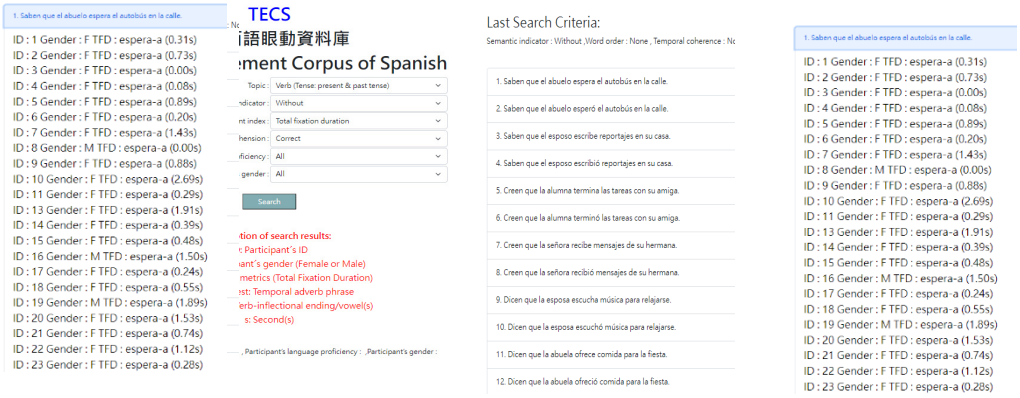

Figure 1. Taiwan Eye-Movement Corpus of Spanish (TECS)

In addition to compiling data from eye-movement indicators in previous research in the TECS corpus, one of TECS' functions is its prediction power. In 2023, we initiated training and model prediction endeavors using the eye movement and subject profile data compiled in the TECS corpus. The objective was to predict learners' comprehension of sentences by analyzing their eye movements during the reading of Spanish sentences. The collected data, including variables such as the subject's proficiency level in Spanish, the stimulus design of the eye movement experiments, and eye movement indicators, represents an immense potential for future developments in machine learning research. By analyzing these variables, we aim to identify patterns that can inform instructional practices. This integration of machine learning aligns with the cognitive load theory (Sweller, 1988), which suggests that managing cognitive load is crucial for effective learning.

For the purpose of training and developing the prediction model, one set of data from an experiment on the processing of tense verbal inflections and temporal adverbs among Taiwanese L3 Spanish learners of different proficiency levels was employed.

Our research involved tracking and detecting learners' eye movement patterns while they read stimuli using eye movement technology. Through an examination of the exported eye movement indicators, we aimed to identify key timing in the process of learning Spanish grammatical points based on learners' input processing. In the analysis of input processing data, we investigated the correlation between verbal inflections (morphological form) and temporal adverbs (lexical items) from the present and simple past tenses using eye total fixation duration time on target areas in a sentence. Our findings revealed the following conclusions: (1) Attention and comprehension: Regardless of verb inflection or temporal adverb, elements appearing earlier received more attention. In the simpler and earlier acquired present tense, when the verb and adverb were temporally incongruent, learners had to pay more attention to understand the sentence's meaning accurately. In the more challenging and later acquired simple past tense, more gaze time was required to focus on the "V-Adv" order than the "Adv-V" order. (2) Lexical items: In both the present tense and simple past tense, regardless of temporal consistency between the adverb and verb, greater attention was directed towards the temporal adverb in the "Adv-V" word order as opposed to the "V-Adv" word order. These experimental results have significant implications for teaching. For example, our findings indicate that learners pay more attention to temporal adverbs than verb inflections in the early stages of learning. This insight can help educators design instructional materials that highlight the critical linguistic elements that learners may otherwise overlook. Additionally, by reducing reliance on temporal adverbs, learners can develop input processing patterns more similar to those of native speakers, thereby facilitating more natural language use. When both the verb and temporal adverb are present in a Spanish sentence, native Chinese speakers tend to rely more on temporal adverbs to interpret the event time. The sooner they reduce their reliance on temporal adverbs for temporal

reference interpretation, the closer their input processing patterns align with those of native Spanish speakers.

Furthermore, we utilized a Machine Learning approach to train a model for eye movement data and predict participants' comprehension accuracy regarding the event time. Our goal is to provide pedagogical insights and enhance learning effectiveness in the long run. Through collaborating with experts in computer science and information engineering, our objective is to achieve cross-domain integration and mutual support by training on eye movement data from Spanish learners and validating prediction models.

2. Literature Review

Sáiz-Manzanares et al. (2021) conducted an analysis using eye-tracking methodology by applying statistical tests and both supervised and unsupervised machine learning techniques. They utilized parameters such as fixations, saccades, blinks, and scan paths. Their eye movement experiment involved collecting personal data, assessing prior knowledge, conducting calibration tests, and making a percentage adjustment. Participants watched a 120-second video and completed a crossword puzzle with five questions related to the video. After data collection, they employed a three-factor fixed effects analysis of variance (ANOVA) to analyze the data, considering factors such as participant type (student vs. teacher), age (over 50 years old vs. under 50 years old), and knowledge level (expert vs. novice). They also used eta squared (η2) for effect size analysis.

Vasseur et al. (2023) examined the methodologies and experimental settings used in eye-tracking research within Information Systems (IS) since 2008. They found that IS research that employed eye-tracking varies in its methodological and theoretical complexity. However, they noted that there are opportunities to better address research questions by leveraging advanced hardware and software options, such as mobile tracking, visualizations, and more sophisticated methodological analysis techniques. Previous research has largely focused on attention-related constructs and used fixation count metrics on desktop computers. Vasseur et al. also recommend that IS researchers combine quantitative and qualitative eye-tracking data with other measurements to triangulate their findings effectively.

Previous studies have explored the use of eye movement data for prediction purposes. For instance, to predict reading ability, Zhan et al. (2016) examined various eye movement indicators, including fixation, saccade, and regression. They found that fixation rate, a sub-indicator of fixation, was negatively correlated with test difficulty, while longer fixation duration indicated deeper and more effortful cognitive processing.

Furthermore, Parikh and Kalva (2020) conducted research to predict potential learning difficulties that may arise from online learning systems and to tailor learning content based on learners' proficiency levels. By observing, measuring, and analyzing learners' reading patterns using an eye movement tracker, this study identified 12 eye-response features that can predict lexical processing and syntactic parsing behavior. The research introduced the non-parametric statistical Feature Weighted Linguistics Classifier (FWLC), achieving an impressive 90% accuracy in predicting learning levels based on individual reading characteristics. The FWLC outperformed other classifiers and demonstrated enhanced accuracy in predicting the difficulty of learning new vocabulary or concepts.

They presented a prediction model that utilized eye movement responses to predict readers' familiarity with vocabulary in contextual paragraphs and the difficulty level of vocabulary learning. The study aimed to enable online learning systems to adapt to individual learning needs. By employing FWLC, the accuracy of the prediction model was examined, revealing a strong correlation between multiple characteristics and learning difficulty. The research concluded that eye responses varied based on different learning difficulties. Additionally, through analysis of eye movement responses, the

researchers identified three relevant features for predicting subjects' vocabulary reading, achieving an accuracy rate of over 90% in predicting vocabulary learning difficulty.

Barral et al. (2021) emphasized the influence of readers' cognitive abilities on cognitive load and reading time. They utilized an eye tracker to predict comprehension and cognition in magazine-style narrative visualizations (MSNVs). This prediction method effectively identified readers' status, enabling individual adjustments to enhance reading benefits.

Random Forest is a classification algorithm based on a Decision Tree, introduced initially in Ho's (1998) "Random Decision Forests" paper. It utilizes binary splitting rules to analyze complex datasets with numerous variables. The algorithm consists of three main steps: (1) dividing the sample data into groups using the Bagging algorithm, generating n training datasets by random sampling with replacement; (2) generating an independent random vector for each training set, randomly selecting m variables for binary splitting attempts without pruning; and (3) integrating the results of the n decision trees generated by the training sets. In this project, which focuses on predicting answer correctness from eye movement data (a binary classification problem), a simple majority vote is employed for integration. The final classification answer is determined by selecting the majority prediction from the n decision trees.

K-Nearest Neighbors (KNN) algorithm, originally introduced by Fix and Hodges in 1951 is a straightforward algorithm that relies on the idea that similar things tend to cluster together. It examines the discrete distribution of data and assigns a category to an input sample based on the classification outcomes of its nearest K neighbors. The algorithm works as follows: calculate the distances between each sample in the dataset, determine the number of neighbors, make a majority decision using the K value, and classify the input sample based on the voting result.

Our research aims to develop a prediction framework for Spanish reading difficulty in Taiwan, as no similar framework exists.

3. Methodology

3.1. Introduce Algorithms in Training Model

To investigate the impact of verbal inflections in different tenses and lexical adverbs on Spanish temporal reference comprehension, we used the Tobii Pro Lab version 1.207 eye tracker and the Eye Tracker Manager version 2.3.7 software to configure the Tobii hardware, which operated at a sampling rate of 250Hz to monitor and record participants' eye movements during the experiment. The experiment was conducted in a controlled environment to minimize external variables. Participants were seated at a fixed distance from the screen, and ambient lighting was maintained at a constant level to prevent glare and reflections.

For data analysis, we defined "Areas of Interest" (AOIs) by highlighting the keywords, such as verb endings expressing tense and temporal adverbs. The eye-tracking metrics, including total fixation duration, were selected in Tobii Pro Lab and exported in Excel format using the "Metrics export" function.

Additionally, we ensured that participants had a consistent measurement of language proficiency levels across various studies by administering the AVANT STAMP 4s exam, which assesses listening, speaking, reading, and writing skills. The STAMP 4S test is a proficiency-based standardized assessment, that evaluates learners' ability to use actual language in real-world contexts, with ratings based on the guidelines issued by the American Council on the Teaching of Foreign Languages (ACTFL). The STAMP 4S rating system includes levels of Novice, Intermediate, Advanced, Superior, and Distinguished, with sublevels of Low, Mid, and High for each level. The STAMP 4S rating scales are aligned with the Common European Framework of Reference for Language (CEFR). The proficiency

levels obtained from the STAMP 4S test proved a consistent reference across different experiments, allowing for reliable comparisons in the prediction model.

We utilized a vector dataset integrating eye movement indicators, comprehension accuracy, and Spanish language proficiency levels. The vectors consisted of eye movement indicators and quantified Spanish proficiency levels, while the vector tagging represented comprehension accuracy. Using a Machine Learning approach, we trained the classification of eye movement indicators, aiming to predict subjects' comprehension accuracy. This analysis aimed to enhance our understanding of participants' learning challenges and improve teaching effectiveness.

We trained a prediction model using the eye movement data based on the integrated results of the eye movement indicators and subjects' Spanish language proficiency. The model was derived, deduced, constructed, and verified using the data collected from an eye movement experiment conducted in 2022. The findings of this study will serve as valuable references for future experiment design and analysis. Figure 2 illustrates the workflow of various stages of this research procedure. We employed machine learning methods and utilized learner eye movement data as features to predict the correctness of their answers.

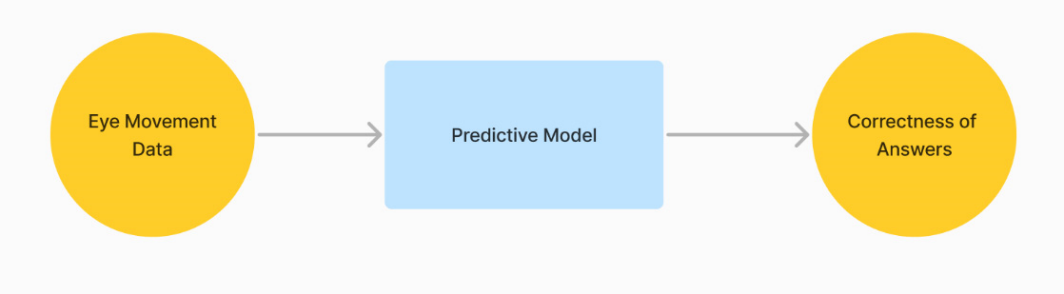

Figure 2. Flowchart Outlining the Prediction Process

In this study, we employed two machine learning methods, Random Forests and the K-Nearest Neighbors (KNN) algorithm, using the Scikit-learn library for Python to predict the correctness of learners' answers. These methods were chosen for their ability to handle complex, nonlinear relationships and their robustness against overfitting.

The Random Forest algorithm combines predictions from multiple decision trees, enhancing prediction accuracy and providing insights into the relationship between learner behavior and language learning outcomes. In contrast, the KNN algorithm determines category assignments based on distance measurements and is known for its simplicity and effectiveness in classification tasks. We evaluated the performance of these prediction models to gain a better understanding of the connection between learner behaviors and language learning outcomes. After evaluation, the Random Forest algorithm was chosen as the final predictive model. It is important to note potential biases and limitations, such as the dataset's inherent bias towards "correct answers," the fixed nature of the experimental environment which may not reflect natural reading conditions, and the reliance on specific eye movement metrics that might not capture all relevant cognitive processes. The flowchart outlining the model training process is presented in Figure 3.

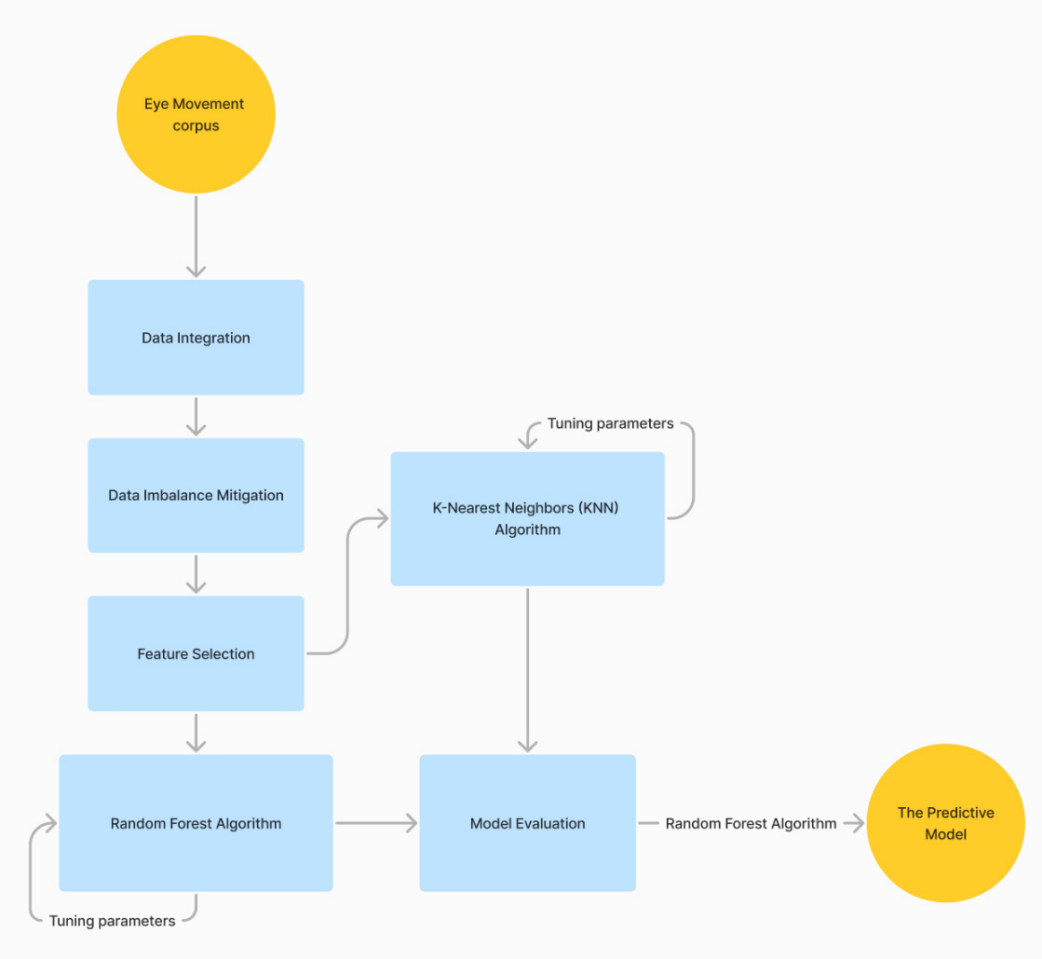

Figure 3. Flowchart for Training the Predictive Model

3.2. Context of Dataset Imbalance

The dataset consisted of 1633 samples, with 1143 samples in the training set and 490 samples in the testing set, following a 7:3 training-to-testing ratio. Additionally, there were no duplicate data in the dataset. It included features such as ID, Gender, Proficiency, Total Fixation Duration/TFD (ADV), TFD (V), Comprehension, and Sentence. The Comprehension feature categorized answers as "answer incorrectly" or "answer correctly."

However, the dataset exhibits an imbalance, with a smaller proportion of instances classified as "answer incorrectly." To mitigate the impact of data imbalance, we utilized the SMOTE (Synthetic Minority Over-sampling Technique) method. This technique involves generating synthetic samples for the minority class by interpolating between existing samples, thus balancing the dataset and improving the model's performance in predicting "answer incorrectly" instances. We used SMOTE to create synthetic samples of "answer incorrectly" instances, which augmented the dataset and achieved a more balanced representation. This rebalancing approach improved the model's accuracy in predicting these instances.

Despite the benefits of using SMOTE, there are potential limitations, including the introduction of noise through synthetic samples and the possibility that these samples may not perfectly represent real-world data distributions. Additionally, relying on a single over-sampling technique could limit the robustness of the findings, indicating a need for future research to explore alternative methods for addressing class imbalance.

4. Results and Discussion

4.1. Performance Evaluation of the Model

This class imbalance presents a challenge and can result in reduced accuracy when predicting "answer incorrectly" data, as indicated by the confusion matrix analysis. The limited representation of "incorrect answers" in our dataset may impact the model's predictive capability in this category, as depicted in Figure 4.

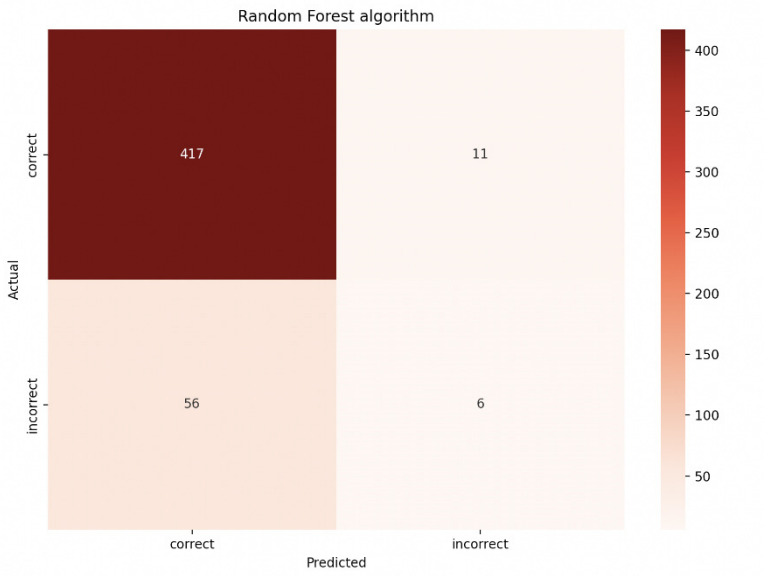

Figure 4. Heatmap Illustrating the Performance of the Random Forest Algorithm Model on an Unbalanced Dataset without SMOTE

The application of the SMOTE method successfully balanced the dataset and improved the model's prediction capability for "answer incorrectly" instances, as depicted in Figure 5. Following the implementation of SMOTE, the Random Forest algorithm model achieved an impressive accuracy rate of 95%, while the KNN algorithm model achieved an accuracy rate exceeding 87%.

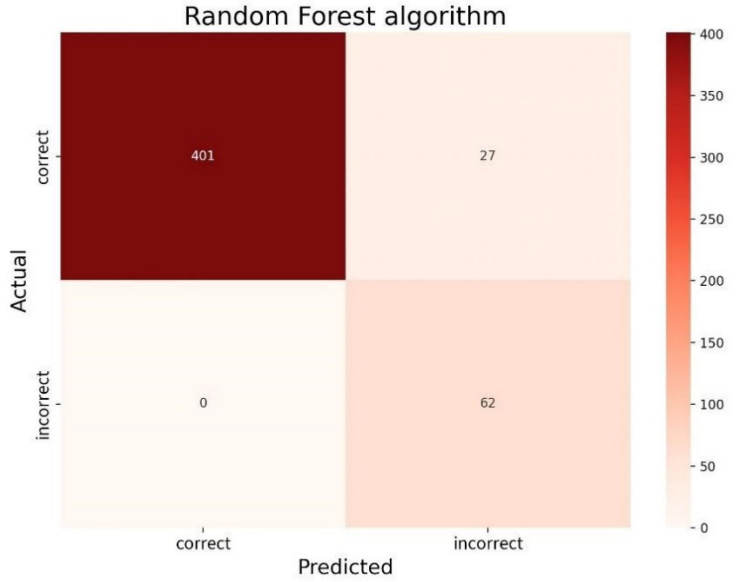

Figure 5. Heatmap Illustrating the Performance of the Random Forest Algorithm Model with SMOTE on a Balanced Dataset

4.2. Random Forest Algorithm: Robust and Versatile for Accurate Predictions

Our model has demonstrated exceptional performance in accurately predicting the correctness of learners' answers, as evidenced by the final accuracy results. This success can be attributed to utilizing the Random Forest algorithm, which highlights its numerous advantages in various aspects.

The Random Forest algorithm's structure enables it to avoid overfitting effectively. With multiple independent decision trees in the ensemble, the diversity among these trees enhances the model's generalization performance. This ensures the model maintains strong predictive power even when presented with unseen data.

Another advantage of the Random Forest algorithm is its ability to address the data imbalance issue in our dataset. In our scenario, there is an imbalance between learners who answer correctly and incorrectly. However, the Random Forest algorithm mitigates the impact of this data imbalance by randomly selecting features and data for building decision trees during each iteration. This adaptability helps maintain the model's predictive power despite the imbalanced dataset.

Moreover, our research leverages the Random Forest algorithm's feature importance evaluation capability. This allows us to assess the influence of each feature on the prediction outcome, providing insights into the key factors in predicting the correctness of learners' answers. In our analysis of feature importance, we discovered that eye movement data features significantly impact the prediction of answer correctness. This finding supports our research hypothesis that learners' eye movements are effective features for predicting answer correctness.

With a remarkable final accuracy of 95%, we can confidently assert that the Random Forest algorithm is a suitable and effective method for this research. This model not only showcases exceptional predictive capabilities but also enhances our understanding of the relationship between data features and the correctness of learners' answers.

While our study demonstrates the effectiveness of using eye movement data to predict comprehension accuracy, several alternative explanations for the observed results should be considered. Factors other than eye movements, such as prior knowledge or contextual understanding, may also play a significant role in learners' ability to comprehend and answer questions correctly. Furthermore, the controlled experimental conditions might not accurately reflect natural reading environments, which could limit the generalizability of our findings.

Additionally, focusing solely on specific eye movement metrics may overlook other relevant cognitive processes involving in language comprehension. Although the use of SMOTE effectively addresses data imbalance, it may introduce noise into the dataset, potentially affecting the model's accuracy. Future research should explore alternative methods for handling class imbalance and incorporate additional cognitive and contextual factors into the predictive models.

5. Implications and Future Research

The findings from our research have significant implications for language teaching practices. By demonstrating that eye movement data can predict the correctness of learners' answers, educators might tailor their instructional strategies to better address students' needs. For example, teachers could use eye-tracking data to pinpoint areas where students struggle and provide targeted interventions to improve comprehension. Additionally, this technology could help develop adaptive learning systems that respond to individual student needs in real time, offering personalized feedback and resources to enhance language acquisition.

This study employed machine learning methods to develop models for predicting learners' language learning outcomes, using data from an eye movement experiment conducted and compiled in TECS in 2022. Given the importance of comprehensible input in second language acquisition,

research on comprehension accuracy and form-meaning mapping in input processing is critical, especially in the initial stage of language learning.

The findings of this study reveal a significant and noteworthy trend: regardless of language learners' proficiency levels, Chinese-speaking learners of Spanish demonstrate a longer total fixation duration (TFD) on temporal adverbs compared to verbal inflections, as illustrated in Figure 6. This suggests that these learners pay more attention to lexical cues (temporal adverbs) than to morphological cues (verbal inflections) when interpreting temporal references during sentence reading. However, as learners advance to the Intermediate High or Pre-advanced levels, the TFD on temporal adverbs declines markedly, as shown in Figure 6. This eye-movement pattern differs from that of Spanish native speakers and learners from morphology-rich languages, such as Romanian (Sagarra & Ellis, 2013). It is plausible that the learners' first language—Chinese, a morphology-poor language—significantly influences their processing strategies for decoding temporal references.

This observation suggests a potential avenue for future research and raises new inquiries regarding how learners from different language backgrounds and proficiency levels allocate their attention when processing various types of input, such as lexical versus morphological cues. The insights gained from this study contribute significantly to a deeper understanding of the cognitive processes involved in language learning and have profound implications for developing more effective language teaching strategies. Thus, there is potential to further utilize the TECS corpus data to predict learner behaviors, which could facilitate future research and enhance teaching practices.

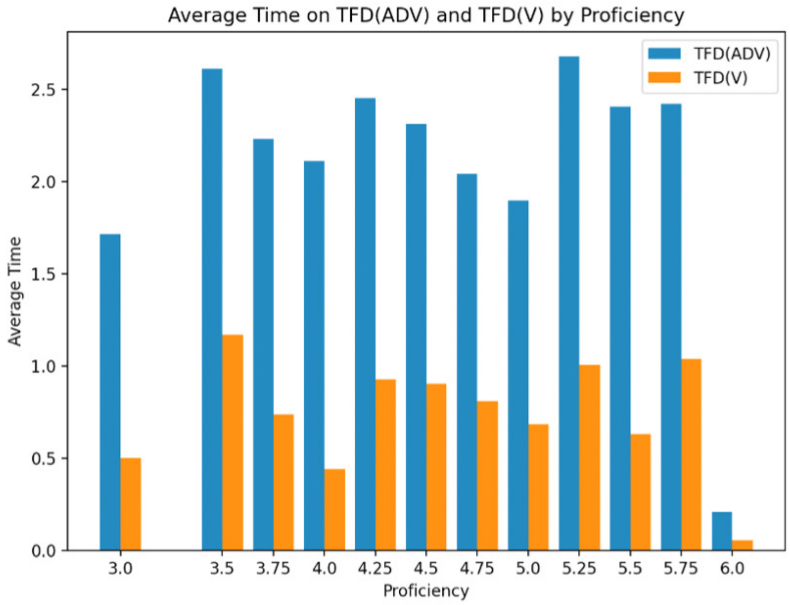

Figure 6. Average Time on TFD (ADV) and TFD (V) by Language Proficiency Level: Request for Additional Information or Data

The findings suggest that specific teaching methods and materials could be developed to enhance learners' understanding of tenses in morphology-rich languages like Spanish. Both temporal adverbs and verbal inflections convey the timeframe of an event or action, and these elements compete for learners' attention during comprehension. Chinese-speaking learners at the Novice and Intermediate levels rely heavily on temporal adverbs rather than verbal inflections. In contrast, native Spanish speakers and learners from morphology-rich languages tend to focus more on verb forms (Sagarra & Ellis, 2013). Therefore, grammar instruction on tenses for Chinese-speaking learners might benefit from a different approach tailored to their proficiency level.

For Novice learners, a pragmatic approach would involve introducing commonly used temporal adverbs that frequently collocate with specific tenses. For example, *ayer* 'yesterday' and *finalmente*

'finally' with the simple past tense, *mientras* 'meanwhile' with the past imperfect, and *mañana* 'tomorrow' with the future and present tenses (Cheng & Lu, 2022). These temporal adverbs can facilitate initial communication; however, over-reliance on lexical cues may reinforce first language transfer and the default strategy of Chinese-speaking learners, potentially hindering their acquisition of verbal tense morphology (Ellis et al., 2014, p. 548).

In addition, given the preference for processing lexical elements over verb forms, VanPatten's processing instruction (1996, 2002) could provide effective grammar teaching. Structured input in processing instruction aims to alter learners' default processing strategies by directing their attention to verb forms, thereby modifying how Chinese-speaking learners engage with input data. For tense learning, the learning materials should include input sentences that lack lexical cues (time adverbs) to temporal reference, thereby compelling learners to focus on and process the morphological cues (verbal inflections) that might otherwise be overlooked.

Furthermore, regarding eye movement data training and model prediction, this study utilized eye movement data collected from eye-tracking experiments focusing on the present and past tenses conducted in 2021 as the training dataset. By analyzing the total fixation duration of learners on specific linguistic elements, the study aimed to predict learners' comprehension accuracy and attention allocation. The research results demonstrate that employing the Random Forest not only successfully established a prediction model with an accuracy of up to 95% but also confirmed the significant impact of learners' eye movement behavior on predicting their comprehension during sentence reading. However, we recognize that the database from one experiment, which included 34 Taiwanese college students, represents a relatively small sample. Additionally, this study employed a limited amount of language material, using only one verb form for each tense and a few temporal adverbs. This selection was intentionally designed to control variables and balance the complexity and salience of the verbal inflections (morphological cues) and time adverbs (lexical cues) at the sentence level. Due to these constraints, we acknowledge that the current eye movement corpus for Spanish learning is in its preliminary stages and insufficient to serve as a robust indicator of language acquisition. Nevertheless, we remain committed to continuously expanding and strengthening the corpus to establish a more solid foundation for future research. For diversity, future research could explore a broader scope of linguistic features; for profusion, future research could recruit a larger sample of learners, expanding the corpus database and testing the prediction model.

As eye-tracking data reflect learners' learning processes and problem-solving strategies instantly, we hope that this study can assist educators and researchers in understanding the relationship between learners' eye movement behaviors on specific vocabulary and their comprehension of the questions. By analyzing how learners' eyes move across particular words, educators can gain insights into how vocabulary understanding impacts overall comprehension. We anticipate integrating these findings into future instructional designs to validate their auxiliary benefits for learning outcomes, ensuring that teaching methods are grounded in solid research about learner behavior. In practical terms, the findings can inform the development of teaching materials and curricula that emphasize critical linguistic elements identified through eye-tracking analysis.

Reflecting on the study's contributions to the field of corpus linguistics and language learning research, this research underscores the value of integrating eye-tracking technology with machine learning algorithms to gain deeper insights into language acquisition processes. The novel approach of using eye movement data to predict comprehension accuracy opens new avenues for exploring the cognitive mechanisms underlying language learning. Moreover, the findings contribute to the growing body of knowledge in corpus linguistics by providing empirical evidence on how learners process linguistic input at a micro-level.

In the future, the eye movement database will also incorporate a more diverse range of research topics and a larger data volume. Additionally, it will further combine artificial intelligence to predict learners' answer correctness and provide tailored learning strategies for individual learners,

enhancing its utility for language teaching and acquisition (da Silva Soares Jr et al., 2023). By expanding the database and incorporating more varied linguistic features, researchers can continue to refine predictive models and enhance our understanding of language acquisition dynamics. Ultimately, this research paves the way for innovative educational technologies and methodologies that leverage eye-tracking data to foster more effective and personalized language learning experiences.

Acknowledgments

This work was supported by the National Science and Technology Council in Taiwan under grant number MOST 111-2410-H-006-020.

References

Barral, O., Lallé, S., Iranpour, A., & Conati, C. (2021). Effect of adaptive guidance and visualization literacy on gaze attentive behaviors and sequential patterns on magazine-style narrative visualizations. *ACM Transactions on Interactive Intelligent Systems (TiiS), 11*(3-4), 1-46.

Cheng, A. C., & Lu, H. C. (2022). A corpus analysis of temporal references and verb tenses cooccurrence in Spanish, English, and Chinese. *Asia Pacific Journal of Corpus Research, 3*(2), 1-16.

da Silva Soares Jr, R., Oku, A. Y. A., Barreto, C. D. S. F., & Sato, J. R. (2023). Exploring the potential of eye tracking on personalized learning and real-time feedback in modern education. *Progress in Brain Research*, *282*, 49-70.

Dussias, P. E. (2010). Uses of eye-tracking data in second language sentence processing research. *Annual Review of Applied Linguistics, 30*, 149-166.

Ellis, N. C., Hafeez, K., Martin, K. I., Chen, L., Boland, J., & Sagarra, N. (2014). An eye-tracking study of learned attention in Second Language Acquisition. *Applied Psycholinguistics, 35*(3), 547-579.

Fix, E., & Hodges, J. L. (1951). *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties*. Texas: USAF School of Aviation Medicine.

Gass, S. M., & Mackey, A. (2007). Input, interaction, and output in second language acquisition. In B. Vanpatten, & J. Williams (Eds.), *Theories in Second Language Acquisition* (pp. 175-200). London: Routledge.

Guan, Y. H. (2020). Exploring the effects of phonological processing on foreign-language reading: An eye-tracking study for Chinese learners of Japanese-as-a-foreign language. *Bulletin of Educational Psychology*, *51*(3), 483-504.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832-844.

Liu, T. S. W., Liu, Y. T., & Chen, C. Y. D. (2019). Meaningfulness is in the eye of the reader: Eye-tracking insights of L2 learners reading e-books and their pedagogical implications. *Interactive Learning Environments*, *27*(2), 181-199.

Parikh, S. S., & Kalva, H. (2020). Feature weighted linguistics classifier for predicting learning difficulty using eye tracking. *ACM Transactions on Applied Perception*, *17*(2), 1-25.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372-422.

Sagarra, N., & Ellis, N. C. (2013). From seeing adverbs to seeing verbal morphology: Language experience and adult acquisition of L2 tense. *Studies in Second Language Acquisition, 35*(2), 261-290.

Sáiz-Manzanares, M. C., Pérez, I. R., Rodríguez, A. A., Arribas, S. R., Almeida, L., & Martin, C. F. (2021). Analysis of the learning process through eye-tracking technology and feature selection techniques. *Applied Sciences*, *11*(13), 6157.

Sweller, J. (1988). Cognitive load during problem-solving: Effects on learning. *Cognitive Science*, *12*(2), 257-285.

VanPatten, B. (1996). *Input Processing and Grammar Instruction in Second Language Acquisition*. New York: Ablex.

VanPatten, B. (2002). Processing instruction: An update. *Language Learning, 52*, 755-804.

VanPatten, B. (2007). Input Processing in Adult Second Language Acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in Second Language Acquisition: An Introduction* (pp.115–135.) Mahwah: Lawrence Erlbaum Associates Publishers.

Vasseur, A., Passalacqua, M., Senecal, S., & Leger, P. (2023). The use of eye-tracking in information systems research: A literature review of the last decade. *AIS Transactions on Human-Computer Interaction, 15*(3), 292-321.

Zhan, Z., Zhang, L., Mei, H., & Fong, P. S. (2016). Online learners' reading ability detection based on eye-tracking sensors. *Sensors*, *16*(9), 1457.

THE AUTHORS

Dr. Hui-Chuan Lu currently serves as a tenured professor within the Department of Foreign Languages and Literature at National Cheng Kung University in Taiwan, specializing in the fields of Spanish Linguistics, corpus linguistics, and syntax studies.
Li-Chi Kao is a part-time researcher at National Cheng Kung University. She will earn her master's degree in computer science and information engineering in 2024.
Zong-Han Li is a part-time researcher at National Cheng Kung University. He will earn his master's degree in computer science and information engineering in 2025.
Dr. Wen-Hsiang Lu is a tenured Professor at National Cheng Kung University. He earned his PhD in information engineering in 2003 at Chiao Tung University, Taiwan. His principal research lies in the field of natual language processing.
Dr. An Chung Cheng is the Chair and professor of the Department of World Languages and Cultures and Director of the Asian Studies Insitute at the University of Toledo in the United States.

THE AUTHORS' ADDRESSES

First and Corresponding Author
Hui-Chuan Lu
Professor
National Cheng Kung University
1 University Road, Tainan, TAIWAN
E-mail: hclu@mail.ncku.edu.tw

Co-authors
Li-Chi Kao
MA student
National Cheng Kung University
1 University Road, Tainan, TAIWAN
E-mail: joycem45@gmail.com

Zong-Han Li
MA student
National Cheng Kung University
1 University Road, Tainan, TAIWAN
E-mail: P76121429@gs.ncku.edu.tw

Wen-Hsiang Lu
Professor
National Cheng Kung University
1 University Road, Tainan, TAIWAN
E-mail: whlu@mail.ncku.edu.tw

An-Chung Cheng
Professor
University of Toledo
2801 W. Bancroft St. Toledo, OH43606, USA
E-mail: anchung.cheng@utoledo.edu

Received: 22 April 2024
Received in Revised Form: 7 August 2024
Accepted: 13 August 2024