IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 May 2024, accepted 20 June 2024, date of publication 24 June 2024, date of current version 2 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3418844

RESEARCH ARTICLE

Readability Grading Based on Multidimensional Linguistics Features for International Chinese Language Education

CHAO ZHANG
College of Foreign Languages, Qufu Normal University, Qufu 273165, China
e-mail: xiangyuezc@hotmail.com

This work was supported in part by the International Chinese Language Education Research Program under Grant 23YH82C, and in part by the Higher Education Youth Innovation Team Project of Shandong Province under Grant 2023RW050.

ABSTRACT Selecting appropriate reading materials for L2 (second language/foreign language) learners is crucial for improving their proficiency in the target language. However, the limitation of effective Chinese text readability classifiers poses a significant hurdle for students and educators in accurately gauging the precise difficulty level of texts in international Chinese education. This research conducted the readability grading of Chinese as a Second Language (CSL) texts by developing a BERT-Based CSL Readability Classifier (BCRC), which utilizes the BERT architecture specifically trained on CSL texts and incorporates multidimensional linguistic features including lexical richness, syntactic complexity and syntax patterns. The model was evaluated using a dataset of CSL texts, and the results indicate that the BCRC model performs effectively in predicting the readability levels of CSL texts. It achieves high mean accuracy of 92.9% across different readability levels, which outperforms baseline classifiers in terms of classification performance, highlighting the enhancement capabilities of multidimensional linguistic features in CSL readability classification models. This study contributes to the field of CSL education by providing a robust readability classifier as a valuable tool for educators, curriculum designers, and developers of CSL learning materials to ensure appropriate text selection based on learners' proficiency levels.

INDEX TERMS Readability grading, multidimensional linguistics features, Chinese as a second language, BERT, BCRC.

I. INTRODUCTION

The increasing recognition and significance attributed to Chinese as a second language (CSL) learning necessitates the implementation of effective pedagogical approaches. Among these approaches, reading emerges as a prominent and essential tool for facilitating second language acquisition and fostering knowledge expansion [1], [2], [3]. According to language learning theories, language acquisition occurs when individuals are exposed to target language materials at a slightly higher difficulty level (i + 1) [4], [5]. Consequently, the need for categorizing reading materials into different levels, known as readability grading, becomes crucial to enable readers to accurately select materials that match their proficiency level. This is equally critical for language teachers who aim to provide learners with reading materials that correspond to their level of difficulty [6]. Therefore, readability grading plays a vital role in the field of international Chinese language education and second language acquisition. By providing a systematic approach to assess the difficulty level of texts, readability grading facilitates the precise selection of reading materials for learners, ensuring that learners engage with texts that appropriately challenge them, leading to more effective language acquisition [7], [8], [9].

Currently, there is a wide range of reading materials available for international Chinese language education [10]. However, the limitation of specialized readability grading tools designed specifically for CSL texts has significantly

The associate editor coordinating the review of this manuscript and approving it for publication was Jad Nasreddine.

limited their effectiveness. Although significant progress has been made in text readability research in fields such as English as a foreign language (EFL), English as a second language (ESL) Russian, and Arabic, the methods developed for these languages are not directly applicable to Chinese due to numerous linguistic differences, such as word frequency and syntax patterns [11], [12], [13], [14], [15], [16], [17]. Moreover, existing readability grading methods, formulations and tools in the field of Chinese language education primarily cater to native Chinese speakers, disregarding the distinctive grammar and learning requirements of CSL learners [18], [19]. Consequently, the current readability grading methods are not suitable for assessing the readability of CSL texts. Nonetheless, the insights gained from readability research conducted in these domains provide invaluable guidance, e.g., the emphasis on multidimensional language features and incorporation of machine learning models into readability grading, making it feasible to develop a readability grading system for CSL texts [20], [21].

The aim of the present study is to develop a readability grading system for CSL reading texts based on multidimensional language features. By considering the unique linguistic characteristics of CSL texts and drawing upon existing research in the field of readability assessment, this system intends to provide accurate and reliable assessments of CSL text readability. The development of such a system will benefit both language learners and teachers by enabling the selection of appropriate reading materials aligned with learners' proficiency levels, thereby enhancing the efficacy of CSL learning.

II. LITERATURE REVIEW

A. TEXT READABILITY GRADING

Text readability grading, also known as readability leveling or difficulty assessment, is a fundamental aspect in various fields, including education, language learning, and content creation [18], [22]. Evaluating the difficulty level of texts allows educators, publishers, and researchers to tailor materials to the needs and abilities of readers, ensuring effective comprehension and engagement. Over the years, the methods employed for text difficulty grading have evolved significantly, transitioning from manual assessments to the development of readability formulas, and more recently, the emergence of machine learning-based readability classification.

Initially, text readability grading relied on human experts who assessed the difficulty of texts based on their subjective judgment and experience [6]. While this approach provided some insights into text readability, it lacked objectivity and scalability. To address these limitations, researchers introduced readability formulas, which aimed to quantify text difficulty using mathematical models. These formulas considered various linguistic and structural features of texts, such as word length, sentence length, and syllable count. The Flesch-Kincaid Grade Level [23], Degrees of Reading Power [24], and Lexile scores [25] are prominent examples of readability formulas that became widely used in the assessment of text difficulty. Despite the advancements brought by readability formulas, readability formulas had their limitations. These formulas primarily focused on surface-level features and did not consider deep linguistic representations of text difficulty (e.g., dependencies, linguistic structure). Additionally, they were developed for specific languages and genres, which restricted their applicability in diverse contexts [26], [27].

With the rapid advancements in natural language processing and machine learning techniques, researchers began exploring the application of these methodologies to the field of text grading [15], [21], [28]. Machine learning-based approaches offered the potential to overcome the limitations of readability formulas and provide more accurate and nuanced assessments. These approaches involved extracting a wide range of linguistic, syntactic, and semantic features from texts and using them as input to train classifiers. Part-of-speech tags, syntactic parse trees, and measures of lexical complexity were among the features employed to capture different dimensions of text difficulty.

More recently, the adoption of deep learning techniques, particularly transformer models such as Bidirectional Encoder Representations from Transformer (BERT), has revolutionized the field of text grading [29], [30], [31]. For instance, Mi et al. introduced a BERT-based classification model, which demonstrated an average accuracy of 85.3%, surpassing that of baseline models [30]. These neural network-based models have the ability to capture linguistic nuances, and intricate relationships between words, contextual information, enabling more sophisticated and context-aware assessments. Leveraging large-scale annotated datasets and pre-trained language models, researchers have achieved significant improvements in the accuracy and scalability of readability classifiers. Attention mechanisms and fine-tuning strategies have further enhanced the performance of these models.

This field continues to advance with researchers exploring novel avenues for text difficulty grading, including incorporating multidimensional features, such as linguistic characteristics, for a more holistic understanding of text readability [28]. Specialization based on domain characteristics is another promising direction (e.g., readability grading for Russian, Arabic, CSL), allowing for tailored assessments that account for factors [7], [12], [15]. In conclusion, the evolution of text difficulty grading methodologies from manual assessments to readability formulas and machine learning-based readability classification has significantly advanced the field. The integration of machine learning techniques has addressed the limitations of traditional methods and opened up new possibilities for more accurate, scalable, and context-aware assessments. By improving our understanding of text difficulty, these advancements contribute to the development of effective educational materials, personalized instruction, and enhanced language learning experiences.

B. CHINESE TEXT CLASSIFICATION AND CSL TEXT READABILITY GRADING

Chinese, a widely used language, has attracted significant research attention in the field of text classification. Numerous studies have been conducted to classify Chinese texts across various domains and genres, such as Chinese CohMetrix [26] and *Common Text Analysis Platform* [32]. These studies have employed diverse methodologies, ranging from traditional machine learning techniques to advanced deep learning approaches, including traditional machine learning algorithms such as Naive Bayes, Support Vector Machine (SVM), and Decision Trees, as well as deep learning architectures like CNNs, RNNs, and Transformer models [32], [33], [34], [35], [36]. It highlights the diverse applications of Chinese text classification, such as sentiment analysis, topic modeling, document classification, and spam detection. The extensive research in Chinese text classification reflects its importance in various applications.

In contrast to the abundance of research on Chinese text classification, the study of difficulty grading in CSL texts is relatively limited. CSL text classification primarily relied on manual categorization initially, where experts subjectively categorized texts based on their linguistic expertise. However, this approach had limitations in terms of objectivity and scalability, leading to the development of CSL readability formulas, providing a more systematic and quantifiable approach to CSL text classification and difficulty grading. These formulas incorporated various linguistic and structural features, such as syntactic complexity, sentence length, and word length [37], [38]. Readability formulas facilitated the assessment of CSL text complexity and aided in the development of appropriate instructional materials. Difficulty grading in CSL involves assessing the complexity and readability of Chinese texts intended for non-native learners. Understanding the difficulty level of CSL texts accurately is crucial for designing appropriate instructional materials and fostering effective language learning experiences. However, due to the unique language features of CSL, such as grammatical points and syntactic features, difficulty grading in CSL requires tailored approaches. Few studies on readability in the CSL field have utilized SVMs. SVMs are supervised machine learning algorithms known for their classification capabilities, which excel at separating data points into distinct classes by identifying hyperplanes that maximize separation in high-dimensional spaces. For example, Sung et al. have presented the Chinese Readability Index Explorer for Chinese as a Foreign Language (CRIE-CFL), a tool specifically developed to categorize texts according to the proficiency levels of language learners. In this study, a support vector machine model was employed, integrating 30 linguistics features extracted from a corpus of 1,578 texts, which were initially classified by expert CFL teachers. The SVM model demonstrated commendable accuracy in accurately predicting the proficiency levels of texts; however, these methods also have other limitations, such as the inability to accurately match the requirements of Chinese language teaching and difficulty in adapting to the various linguistic features and dynamic changes in the language for relying on explicit linguistic features.

Chinese text classification techniques have made significant progress, with various applications emerging. However, there is currently limited research on applying multidimensional models to CSL text difficulty classification. CSL text classification has evolved from manual categorization to readability formulas and advanced to machine learning-based approaches. Given that difficulty grading in CSL is essentially a text classification, the integration of machine learning techniques will greatly improve the objectivity, scalability, and accuracy of CSL text classification, offering valuable insights for the research community and empowers more effective CSL education. Some machine learning model using in prior studies, e.g., SVMs, have some limitations in CSL text readability classifications. SVMs treat each data point independently and do not consider the contextual relationships between words or sentences and rely on pre-defined features, such as n-grams or word frequencies, which may not capture the full complexity of natural language. This can limit the ability to capture nuanced semantic information present in the text. In light of the aforementioned considerations, we put forth a proposal to design and assess a BERT-based classifier for CSL readability classification. BERT has exhibited exceptional capabilities in diverse natural language processing tasks, including text classification. Compared with SVMs, BERT is a deep neural network that learns representations directly from the text data, allowing it to capture intricate linguistic patterns and relationships and have the capability to acquire implicit yet comprehensive semantic information from CSL text data. By harnessing the contextualized word embeddings furnished by BERT, we can adeptly capture the semantic and syntactic attributes inherent in CSL texts. The proposed classifier underwent training on an extensive corpus of CSL texts, integrating a range of linguistically informed features that pertain specifically to the complexities of CSL text readability. Subsequently, we conducted a comprehensive evaluation of the classifier's performance, employing well-established evaluation metrics. Furthermore, we compared the performance of the BERT-based classifier against a baseline model as a means to showcase its efficacy in classifying CSL texts. Our ultimate objective is to forge an accurate, dependable, and efficacious CSL readability classifier that can be fruitfully employed within educational contexts, thereby facilitating the judicious selection and adaptation of teaching and learning materials tailored to varying levels of proficiency. This research endeavor stands to contribute to the advancement of CSL education and proffer valuable insights into the field of text classification.

III. DESIGN AND EVALUATION OF BERT-BASED CSL READABILITY CLASSIFIER (BCRC) WITH MULTIDIMENSIONAL LANGUAGE FEATURES

Considering the exceptional performance achieved by integrating linguistic features with BERT in text classification,

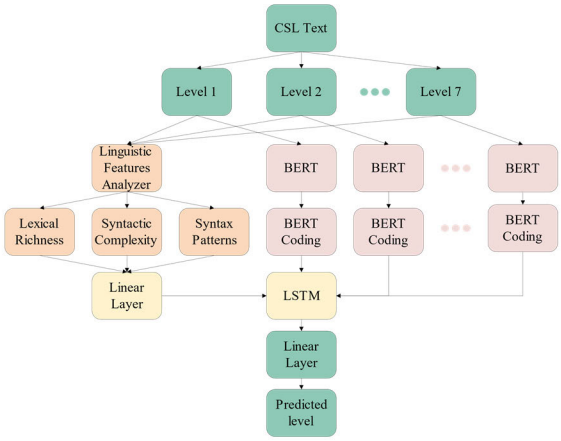

FIGURE 1. The architecture of BCRC.

we have developed the BERT-Based CSL Readability Classifier (BCRC) (Figure 1).

This approach employs the Linguistic Feature Analyzer to meticulously analyze CSL texts across various proficiency levels, thereby capturing multidimensional language features. By harnessing the power of BERT, the BCRC skillfully extracts deep contextual language features from the texts, enabling a comprehensive understanding of the underlying linguistic dynamics. The BCRC seamlessly integrates these extracted features to deliver precise predictions regarding the text's readability level.

A. DATA COLLECTION AND PREPROCESSING

The first step in designing the BERT-based CSL readability classifier (BCRC) is to collect a dataset of CSL texts in English at different readability levels. To obtain a comprehensive dataset of CSL texts, a collection procedure is followed. We gathered a diverse range of CSL texts from various textbook sources, including *Boya Chinese Elementary, Developing Chinese, HSK Standard Course, and Road to Success*, etc. These textbooks are authoritative materials and are widely used in the field of CSL education in Asia, Europe, America, Africa, etc., including labels indicating the readability level of each text, covering a wide spectrum of readability levels and topics commonly encountered by CSL learners. These readability levels are aligned with the Chinese Proficiency Grading Standards for International Chinese Language Education (CPGS), which was approved by the National Language Commission and published by the Ministry of Education, PRC., and was the authoritative international standards for Chinese language education proficiency levels [39].

Preprocessing of the dataset involves tokenizing the texts into sentence and discourse units and applying necessary cleaning or normalization techniques. The texts will be tagged with multidimensional language features (e.g., lexical complexity, syntactic patterns), which were derived through linguistic analysis based on existing readability scales.

B. PRETRAINING BERT ON CSL CORPUS

To ensure that the BCRC is capable of capturing the linguistic intricacies specific to CSL texts, we performed a comprehensive pretraining process using the large-scale CSL corpus. During the pretraining process, the BERT model underwent rigorous training on a large amount of unlabeled CSL text data. The primary objective of the pretraining stage was to equip the BERT model with a deep understanding of the linguistic patterns, syntactical structures, and lexical nuances prevalent in CSL texts. By exposing the model to this extensive corpus, it could effectively learn and encode the complex relationships between words, characters, and sentences in CSL. Specifically, we employed two fundamental tasks: masked language modeling (MLM) and next sentence prediction (NSP).

By jointly training the BERT model on these two tasks using the CSL corpus, we effectively fine-tuned the model's language representation capabilities, enabling it to capture the specific linguistic nuances and patterns inherent in CSL texts. Through this thorough pretraining process, the BERT model developed a deep understanding of CSL language structures, including character complexity, sentence syntax, and discourse. This equipped the model with the necessary linguistic knowledge to better encode and interpret CSL texts, forming the foundation for the subsequent stages of BCRC development

C. FINE-TUNING BERT WITH CSL READABILITY DATA

After the pretraining stage, we further fine-tuned the BERT model using a CSL readability dataset that consisted of CSL texts labeled with corresponding readability levels. This fine-tuning process aimed to optimize the model's ability to predict the readability levels of unseen CSL texts accurately.

To accomplish this, we employed a supervised learning approach, leveraging the labeled CSL readability dataset. The dataset was divided into training, validation, and testing sets. The training set was used to update the BERT model's parameters, while the validation set was used to monitor the model's performance and perform hyperparameter tuning. The testing set remained untouched until the final evaluation stage.

During the fine-tuning process, we utilized the MLM task to train the model on the CSL readability data. This task allowed the model to learn the specific readability-related linguistic patterns present in CSL texts. Additionally, we employed a specialized loss function to optimize the model's predictions for readability levels. This loss function measured the discrepancy between the predicted readability levels and the ground truth labels. By minimizing this discrepancy, the model was encouraged to make more accurate readability level predictions.

Throughout the fine-tuning process, we iteratively updated the model's parameters using gradient-based optimization algorithms, such as stochastic gradient descent (SGD) or Adam optimizer. We carefully tuned hyperparameters, including learning rate, batch size, and regularization

techniques, to ensure optimal model performance and prevent overfitting.

To assess the effectiveness of the fine-tuned BCRC, we evaluated its performance on the testing set, which contained unseen CSL texts. We employed various evaluation metrics, including accuracy, precision, recall, and F1 score, to measure the model's ability to predict the correct readability levels. The fine-tuning of the BERT model with CSL readability data aimed to enhance the model's understanding of the specific factors that contribute to the readability of CSL texts. By leveraging the labeled CSL readability dataset, the model learned to generalize potential knowledge from the training set to make accurate predictions on unseen CSL texts, effectively creating a BERT-based CSL Readability Classifier capable of assessing the readability levels of CSL texts with high accuracy and reliability.

D. MULTIDIMENSIONAL LANGUAGE FEATURES AND FUSION IN BCRC

In addition to leveraging the powerful language representation learned by BERT, we incorporated multidimensional language features to enhance the accuracy and effectiveness of the BERT-based CSL Readability Classifier (BCRC). These features aimed to capture the diverse linguistic characteristics and complexities inherent in second language texts, providing a more comprehensive understanding of their readability levels. Vocabulary and syntax play a significant role in the comprehension and readability of texts [40], [41]. To evaluate the readability of the CSL texts, we integrated lexical and syntactic features into the BCRC. By considering these attributes, the BCRC can better assess the impact of vocabulary difficulty on the overall readability of CSL texts. The following multidimensional language features were verified in high validity in Chinese text classification and were incorporated in BCRC.

1) LEXICAL RICHNESS

Lexical richness, as a multidimensional construct, encompasses various sub-dimensions, namely lexical variation, lexical complexity, lexical density, and lexical errors [42]. Notably, researchers have observed notable distinctions in the relationship between diverse lexical and syntactic indicators and the quality of writing performance [43]. To assess lexical richness in CSL text, three indicators are primarily used: lexical variation, lexical complexity, and lexical density. These indicators were applied to the data used in this study, which was extracted from authoritative textbooks. Lexical variation refers to the range of vocabulary used by learners. In the field of language acquisition research, measurement methods based on type and token have been widely adopted for assessing lexical variation in English as a second language due to their high validity and practicality (Laufer & Nation, 1995). Type refers to all the distinct words in a text, while token refers to the total number of words. The type-token ratio (TTR) is an important indicator for measuring lexical variation [44]. However, TTR has been criticized for its susceptibility to text length, as it tends to decrease with longer texts. In response to this, several modified TTR indicators, such as LogTTR and Root TTR (RTTR), have been proposed by scholars. Lu compared around 20 different measurement methods of lexical variation and found that TTR and its modified versions remained the most reliable indicators. Therefore, this study adopts the RTTR, which is computed as type/sqrt(token), where type represents the number of unique vocabularies, and token denotes the total number of words. This metric provides an indication of the lexical diversity within the text.

Lexical density, as defined by Ure, pertains to the ratio of content words to the overall word count within a given text [45]. In the context of Chinese language, content words encompass nouns, verbs, adjectives, adverbs, classifiers, numerals, pronouns, interjections, and onomatopoeic words. The computation formula for lexical density (type density) is expressed as follows: Lexical density (type density) = total number of unique content word types / total number of unique word types.

Lexical sophistication measures the proportion of low-frequency words in a text [42]. In this study, CSL vocabulary is categorized into 5 levels based on the *Chinese Proficiency Vocabulary and Chinese Character Level Outline* (2001). The levels range from A to E, with A and B representing high-frequency words and C and D representing low-frequency words. This study focuses on the analysis of lexical sophistication in written Chinese as a second language by examining the proportions of A and D-level words in the text. This method has been proven to be effective in Chinese as a second language research [3], [10]. The calculation formula for lexical sophistication (low-frequency word ratio) is: Lexical sophistication (low-frequency word ratio) = sum of C and D-level word types / total number of word types.

2) SYNTACTIC COMPLEXITY

Syntactic complexity plays a crucial role in determining the readability of CSL texts. To capture these structural characteristics, the syntactic complexity feature was incorporated into the BCRC. These features enable the model to assess the impact of sentence structures on the overall readability of CSL texts, which is particularly important for language learners.

Syntactic complexity is a crucial aspect of language analysis [46]. Drawing on previous research on syntactic complexity in Chinese texts, this study employs the reliable and validated measure of sentence length to assess syntactic complexity [3]. Sentence length refers to the average number of words per sentence in a given text. In this study, the syntactic complexity of Chinese as a second language is computed using the following formula: Syntactic complexity (measured by sentence length) = total number of words / total number of sentences. This formula allows for a quantitative evaluation

of the syntactic complexity present in the analyzed Chinese as a second language texts.

3) SYNTACTIC PATTERN

Syntactic point refers to the fundamental syntactic concepts or knowledge that language learners need to acquire at different proficiency levels throughout the stages of their language acquisition process. These points include but are not limited to noun phrases, compound sentences, and other syntactic structures. The mastery of syntactic points is directly related to learners' language proficiency and significantly influences their reading comprehension of texts.

In order to analyze the syntactic complexity of the articles, the researchers utilize the natural language processing tool, Stanford CoreNLP [47]. CoreNLP allows for a comprehensive preprocessing of the articles by providing various linguistic annotations, including word segmentation, part-of-speech tagging, syntactic tree analysis, and dependency parsing. These annotations provide valuable insights into the structural aspects of the text, enabling a deeper analysis of the syntactic feature (i.e., syntactic pattern) present in the data.

To effectively detect and extract particular syntactic patterns of interest, the implementation of regular expressions was employed to match such patterns. The primary objective of the BCRC is to systematically capture and retrieve sentences that encompass specific grammatical structures or syntactic phenomena. Through the meticulous construction of syntactic patterns using regular expressions, the BCRC endeavors to precisely target the desired syntactic focal points. Subsequently, the BCRC is deployed to search for and retrieve all instances that conform to the specified syntactic patterns. This retrieval process enables a comprehensive identification and analysis of sentences that exhibit the intended syntactic elements, thereby facilitating a deeper comprehension of the syntactic intricacies within the corpus. As highlighted by Wolfe-Quintero et al., the ratio has been established as a reliable indicator for evaluating written texts [44]. In this study, the syntactic pattern ratio, quantified as the number of syntactic patterns per word, was utilized as a robust quantitative descriptor to assess the prevalence of specific syntactic patterns within the CSL corpus.

To integrate the BERT-based contextual embeddings with complexity and syntactic patterns, we applied Long Short-Term Memory (LSTM), which captures long-term dependencies and relationships in sequential data, such as natural language text [48]. The complexity and syntactic patterns were combined through a linear layer, and the resulting encoding, along with the BERT embeddings, was passed into the LSTM layer for further processing. Finally, the combined representations were processed by another linear layer. The fused representations capture both the contextual information from BERT and the linguistic features related to complexity and syntax. In this approach, each feature type (BERT embeddings, complexity points, and syntactic patterns) is multiplied by a respective weight. The weighted feature vectors are then summed to obtain the fused representation, in which the weights assigned to each feature type can be adjusted to balance their contributions and optimize the classifier's performance. The vector fusion phase plays a vital role in BCRC with integrated language features. This integration allows the classifier to leverage both types of information in predicting the readability of CSL texts.

The incorporation of multidimensional language features in the BCRC acknowledges the multifaceted nature of CSL texts and ensures that the model considers various linguistic factors that influence their readability. By combining these features with the powerful language representation capabilities of BERT, the BCRC will demonstrate a more comprehensive and holistic approach to CSL readability assessment. This enriched representation allows the model to capture the intricate linguistic nuances and complexities specific to CSL, ultimately improving the accuracy and precision of readability level predictions.

IV. TRAINING AND EVALUATION

To develop a robust and accurate BERT-based CSL Readability Classifier (BCRC), we conducted rigorous training and evaluation procedures. These steps aimed to optimize the model's performance, assess its effectiveness, and ensure reliable readability predictions for CSL texts.

A. TRAINING

The training process involved feeding the labeled CSL readability dataset, consisting of 18430 CSL texts with assigned readability levels. Additionally, the texts were annotated with complexity and syntactic patterns, capturing linguistic features related to text readability. We employed the combined representation of BERT embeddings and multidimensional language features to train the model. The training set was used to update the BCRC's parameters through backpropagation and gradient descent optimization algorithms.

During training, the models were iteratively adjusted to minimize the discrepancy between the predicted readability levels and the ground truth labels. This process involved computing the loss function, which quantified the difference between the predicted and actual readability levels. Through gradient-based optimization, the BCRC learned to make more accurate predictions and adapt its parameters to the specific characteristics of CSL texts. Furthermore, hyperparameter tuning was performed to optimize the model's performance. We explored various hyperparameters, including learning rate, batch size and epochs, to ensure the BCRC achieved optimal accuracy and generalization on unseen CSL texts. The hyperparameter configuration were carefully selected through extensive empirical evaluation on the test dataset. As detailed in Table 1, the BCRC model was trained for 10 epochs, a duration deemed sufficient to facilitate convergence based on the observed learning dynamics. The learning rate, a critical hyperparameter governing the optimization process, was set to 5e-6, a value identified through rigorous experimentation to yield stable and effective updates during

TABLE 1. Hyperparameter of the baseline model and BCRC model.

Hyperparameter	Baseline Model	BCRC
learning rate	2e-5	5e-6
batch size	32	16
epochs	10	10
hidden size	-	128
classes	7	7
dropout	0.2	-
features	3	3

training. Additionally, the model was trained in a batch-wise fashion, with each update computed over 16 samples. This batch size was empirically determined to strike an appropriate balance between the introduction of stochasticity, which can aid the model in escaping suboptimal local minima, and computational efficiency. Furthermore, the model incorporated a customized hidden size of 128, a design choice that may have enhanced the model's representational capacity and ability to capture intricate patterns within the 7-class dataset. Notably, the model architecture did not explicitly incorporate dropout, a common regularization technique employed to mitigate overfitting. The decision to forgo dropout was likely informed by the model's performance on the validation set during the hyperparameter tuning process. Additionally, through experimental evaluation, the final model was configured to utilize 3 linguistic features, namely lexical variation, lexical proportion of A-level word, and lexical proportion of D-level word, as these features were found to yield the optimal results for the 7-class task. Through the careful selection of these key training hyperparameters, the models were optimized to deliver robust and reliable performance on the 7-class readability classification task at hand.

B. EVALUATION

After training, we evaluated the performance of the BCRC using the testing set, which contained 950 unseen CSL texts with about 30000 words, ranging from beginner to advanced readability levels. The testing dataset is derived from authoritative international Chinese language textbooks. The evaluation aimed to assess the BCRC's ability to accurately predict the readability levels of CSL texts.

To assess the effectiveness of BCRC in classifying CSL text readability, we compare its performance with the baseline model (i.e., the readability classifier using only BERT embeddings). Matplotlib was utilized to generate a scatter plot with reduced dimensions, effectively showcasing the model's performance [49]. In this plot, data points depicted in red signify instances of mismatching data, indicating incorrect classifications. Conversely, green color is assigned to data points representing matching data, signifying correct classifications. We utilized various evaluation metrics to measure the effectiveness of the BCRC, including accuracy, precision, recall, and F1 score. Accuracy represents the overall correctness of the predicted readability levels, which is calculated by averaging the accuracy of predictions for all classes. Specifically, it computes the ratio of the number of correct predictions for each class to the total number of samples and then takes the average of these ratios. Precision measures the proportion of correctly predicted positive instances (readability levels) out of all predicted positive instances. Recall, on the other hand, quantifies the proportion of correctly predicted positive instances out of all actual positive instances. The F1 score combines precision and recall to provide a balanced measure of the model's performance. Scikit-learn library [50], a popular and authorized machine learning library in Python that provides various tools and algorithms for tasks, was used to automatically calculate evaluation metrics to ensure the validity. This comparison allows us to assess the effectiveness of our BCRC in predicting the readability levels of CSL texts. By analyzing these evaluation metrics, we gained insights into the BCRC's effectiveness in predicting CSL text readability levels. Comparisons with traditional readability formulas and other state-of-the-art classifiers were also conducted to validate the superiority of the BCRC in accurately assessing the readability of CSL texts.

The training and evaluation processes ensured that the developed BCRC achieved high performance and reliability in predicting the readability levels of CSL texts. Through careful training and rigorous evaluation, the BCRC demonstrated its effectiveness in capturing the complexities of CSL texts and providing valuable insights for CSL educators and learners. In summary, the design of the BCRC with language feature integration involves data collection and preprocessing, BERT-based representation learning, complexity and syntactic pattern extraction, vector fusion of BERT embeddings with linguistic features, readability classification, and evaluation. This design allows for the integration of both contextualized embeddings from BERT and linguistic information related to complexity and syntactic patterns, enabling a more comprehensive and accurate prediction of CSL text readability.

V. RESULTS

In this section, we present the results of Baseline Model and our BCRC with integrated language features. The classifier was trained on a dataset of CSL texts in English, which were annotated with readability levels and enriched with complexity and syntactic patterns. We evaluated the performance of the classifier using various evaluation metrics and compared it with baseline models.

A. BASELINE MODEL

As the baseline model underwent evaluation for its proficiency in predicting the readability levels of texts, it showcased distinct levels of accuracy across varied difficulty levels. The scatter plot (Figure 2) unveils a sparse presence of red points, indicating the baseline model's exceptional efficacy in readability grading. The classifier attained an aggregate accuracy of 0.846 and an F1 score of 0.677 in

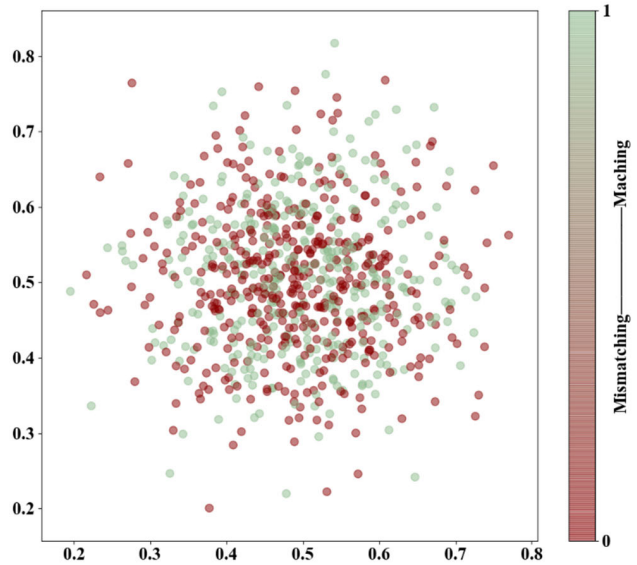

FIGURE 2. The scatter plot depicting the performance of the baseline model in readability grading.

TABLE 2. Performance metrics of the baseline model.

Beginner	Level	Accuracy	Precision	Recall	F1
	1	0.954	0.859	0.945	0.895
Beginner	*2*	0.838	0.815	0.713	0.742
	3	0.831	0.587	0.570	0.577
	4	0.880	0.626	0.639	0.632
Intermediate	*5*	0.787	0.577	0.578	0.578
	6	0.820	0.611	0.633	0.620
Advanced	*7*	0.811	0.678	0.722	0.694
	M	0.846	0.679	0.686	0.677

the prediction of CSL text readability. Detailed precision, recall, and F1 scores for individual readability level classes are presented in Table 2.

For the beginner level, the BERT model achieved an accuracy of 0.831 to 0.954, precision of 0.587 to 0.859, recall of 0.570 to 0.945, and F1 score of 0.577 to 0.895. These results indicate that the model accurately classified beginner-level texts, with an imbalance between different levels.

For the intermediate level, the BERT model achieved an accuracy of 0.787 to 0.880, precision of 0.577 to 0.626, recall of 0.578 to 0.639, and F1 score of 0.578 to 0.632. These metrics suggest that the model performed well in classifying intermediate-level texts, with higher accuracy.

For the advanced level, the BERT model achieved an accuracy of 0.811, precision of 0.678, recall of 0.722, and F1 score of 0.694. These results indicate that the model demonstrated good accuracy and a balanced performance in classifying advanced-level texts

B. BCRC MODEL

The BCRC model also underwent evaluation for its performance in predicting the readability levels of texts across

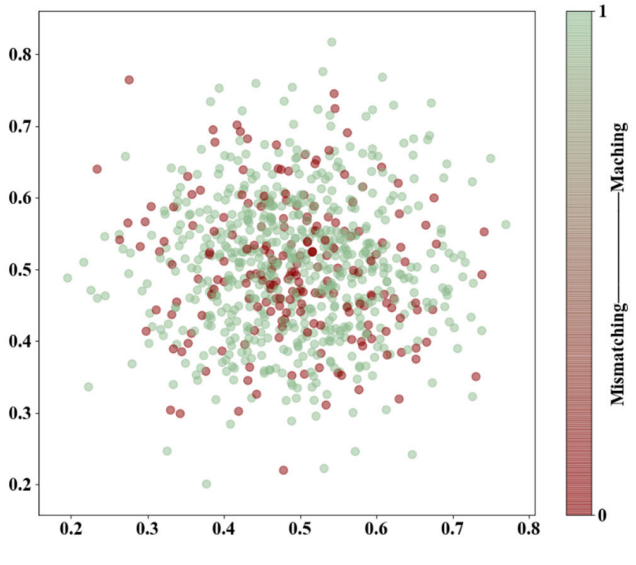

FIGURE 3. The scatter plot depicting the performance of the baseline model in readability grading.

TABLE 3. Performance metrics of the BCRC.

Beginner	Level	Accuracy	Precision	Recall	F1
	1	0.976	0.920	0.967	0.942
Beginner	*2*	0.911	0.882	0.876	0.879
	3	0.927	0.841	0.818	0.829
	4	0.935	0.784	0.859	0.815
Intermediate	*5*	0.914	0.854	0.775	0.807
	6	0.945	0.930	0.803	0.852
Advanced	*7*	0.895	0.801	0.874	0.830
	M	0.929	0.859	0.853	0.851

different readability levels. In comparison to the performance of the baseline model, the BCRC model demonstrates a superior level of performance. The scatter plot (Figure 2) visually presents a diminished and more scattered distribution of red data points in contrast to Figure 1, thereby indicating the heightened capability of BCRC in assessing the readability of CSL texts. It consistently outperformed the baseline model, demonstrating higher accuracy (M = 0.929), precision (M = 0.859), recall (M = 0.853), and F1 scores (M = 0.851) (Table 3).

For the beginner level, the BCRC model achieved an accuracy of 0.911 to 0.976, precision of 0.841 to 0.920, recall of 0.818 to 0.967, and F1 score of 0.829 to 0.942. These metrics indicate that the BCRC model classified beginner-level texts with high accuracy, precision, and recall.

At the intermediate level, the BCRC model achieved an accuracy of 0.914 to 0.945, precision of 0.784 to 0.930, recall of 0.775 to 0.859, and F1 score of 0.807 to 0.852. These results highlight the BCRC model's strong performance in accurately classifying intermediate-level texts, with high precision and recall scores.

For the advanced level, the BCRC model achieved an accuracy of 0.895, precision of 0.801, recall of 0.874, and

F1 score of 0.830 These metrics indicate that the BCRC model exhibited excellent performance in accurately classifying advanced-level texts, with high precision and recall scores.

Both the baseline model and BCRC models demonstrated the capability to predict the readability levels of texts. However, a comparison reveals that the BCRC model consistently outperformed the BERT model across all levels. The BCRC model achieved higher accuracy, precision, recall, and F1 scores than the baseline model, indicating a strong ability to correctly classify CSL texts. These results highlight the effectiveness of integrating BERT embeddings with linguistic features in improving the readability classification of CSL text. These results indicate that the BCRC model has a stronger classification performance in accurately categorizing texts into their respective readability levels.

In summary, the BCRC model exhibited superior performance compared to the BERT model in predicting the readability levels of texts. The BCRC model consistently achieved higher accuracy, precision, recall, and F1 scores across all difficulty levels, indicating its stronger classification ability. These findings highlight the potential of the BCRC model in assisting educational content recommendation and language processing applications. Further research and evaluation on larger and more diverse datasets could further validate and generalize these findings.

VI. DISCUSSION

The results of our experiment revealed that the BCRC exhibited superior performance in accurately assessing the readability levels of CSL texts when compared to the baseline model. This finding underscores the significance of incorporating multidimensional language features to enhance the performance of readability assessment tools.

One key aspect that greatly contributed to the remarkable performance of BCRC was the inclusion of multidimensional language features. CSL texts exhibit distinctive linguistic characteristics that set them apart from native Chinese texts. These linguistic variations encompass factors such as lexical density, sentence length, and syntax patterns, which are critical in determining the readability and comprehensibility of CSL texts [6], [51]. Acknowledging the importance of these particular linguistic features, BCRC integrated them into readability assessment framework, leading to enhanced accuracy and reliability in assessing the readability of CSL texts. This aligns with the work of Sung et al., who incorporated linguistic features into a CSL text classifier [7], achieving an impressive accuracy of 89.71% in CSL readability grading. This emphasis on tailoring readability assessment tools to the specific linguistic properties of the target language is of paramount importance. The unique linguistic characteristics of CSL texts demand a nuanced approach to readability assessment. Traditional readability formulas and generic classifiers generally fail to account for the intricacies of CSL-specific language features. By specifically incorporating these features into its assessment model, BCRC demonstrated a deep understanding of the linguistic nuances present in CSL texts, thereby enhancing the accuracy and reliability of its readability assessments. For instance, CSL texts frequently employ vocabulary that is specific to the domain of second language acquisition, encompassing terms related to language learning, cultural contexts, and language use, which affects the lexical characteristics in terms of variation and complexity. BCRC's inclusion of these domain-specific vocabulary features enabled it to capture the complexity of CSL texts more effectively, resulting in assessments that aligned more closely with the intended proficiency levels of CSL learners. Additionally, CSL texts often display distinctive syntactic structures and patterns. Incorporating these syntactic features into BCRC model allowed for a more nuanced analysis of text readability. By considering these diverse and unique sentence structures, word order variations, and sentence-level coherence in CSL texts, BCRC was able to provide more accurate and tailored readability assessments, ensuring that the classification reflected the actual difficulty level experienced by CSL learners.

Another factor that significantly contributed to the superior overall performance of BCRC was the integration of the BERT model. The seamless integration of BERT's contextual understanding and the consideration of domain-specific features enables BCRC to capture the nuances of text complexity in CSL texts, resulting in reliable predictions and accurate assessments across different proficiency levels. BERT's remarkable ability to capture contextual information plays a crucial role in enhancing BCRC's readability assessment capabilities. By leveraging the contextual knowledge provided by BERT, BCRC was able to gain a more nuanced understanding of the text, enabling a more accurate assessment of readability levels in CSL texts. BERT's bidirectional nature allows it to consider both preceding and succeeding words in a sentence, capturing the dependencies and relationships between words. This comprehensive analysis of the context allows BCRC to better gauge the complexity and difficulty of CSL texts, resulting in more precise predictions of readability levels. By incorporating BERT, BCRC goes beyond traditional methods that rely solely on static linguistic features and embraces the context-dependent nature of CSL texts.

The experiment not only assessed the overall performance of BCRC but also focused on predicting readability across different levels. The results demonstrated that BCRC exhibited remarkable accurate and fine-grained classification in CSL, including beginner, intermediate, and advanced levels, which underscores the robustness of BCRC in accurately assessing the readability levels of CSL texts across various proficiency levels. The successful prediction of readability levels by BCRC might attribute to the use of domain- specific design and the BERT mode, which played a significant role in enhancing BCRC's prediction accuracy across different readability levels. Domain-specific design allows the incorporation of knowledge and features that are specifically tailored to the target domain [7], [18]. This includes

leveraging domain-specific linguistic patterns, deep semantics, and contextual information that may be unique to the domain. By exploiting these domain-specific cues, the classifier can better understand and capture the nuances and intricacies of the CSL domain, leading to improved classification accuracy. Domain-specific design also allows for the integration of domain-specific resources, such as the corpus used in this research. These resources can provide additional implicit context information that can aid in the classification process. By incorporating domain-specific knowledge, the classifier can leverage the rich semantic relationships and domain-specific semantic information, resulting in more accurate and fine-grained classification in CSL. Moreover, BERT captures contextual information and generate word embeddings that reflect the meaning and relationships within a text offers a valuable advantage in accurately assessing the complexity of CSL texts. By leveraging the contextual knowledge provided by the BERT model, BCRC can better understand the intricate nuances (i.e., domain-specific linguistic features) of CSL texts and make more precise predictions of readability levels across various proficiency levels. The incorporation of multidimensional language features and the utilization of the BERT model results in BCRC's ability to accurately predict readability levels in CSL texts, catering to readers with different levels of proficiency. This capability is crucial in educational and language learning contexts, as it enables educators and curriculum developers to select appropriate reading materials that match the proficiency levels of CSL learners, facilitating their language development and reading comprehension. Despite of the notable performance of BCRC in predicting readability levels across different proficiency levels, further research and refinements are still possible. Ongoing advancements in deep learning techniques, such as fine-tuning BERT on domain-specific datasets or exploring other contextual models, may provide even more enhanced capabilities for BCRC in assessing readability levels in CSL texts.

VII. CONCLUSION

This research has explored the grading of Chinese as a Second Language (CSL) reading texts based on language features, by developing the BERT-based CSL readability classifier and evaluating its performance of CLS text readability classification. The BCRC model demonstrated excellent overall performance in predicting the readability levels of CSL texts. It achieved high accuracy (0.929), precision (0.859), recall (0.853), and F1 score (0.851), indicating its capability to accurately classify texts into appropriate readability levels. The results revealed the effectiveness and the superior performance of BCRC in CSL text readability grading across different readability levels, shedding light on the effectiveness of the BCRC model in grading the readability of CSL texts, showcasing its strong performance across different readability levels.

This study has yielded significant insights that can be categorized into theoretical, practical, and methodological implications. Theoretical insights from this research provide empirical evidence supporting the important influence of vocabulary and syntactic features on text readability. The findings confirm the notion that the selection and arrangement of words and sentence structures significantly impact the readability of CSL texts. This theoretical insight contributes to the understanding of text readability in CSL education and highlights the importance of considering linguistic features when assessing and selecting reading materials for CSL learners. The practical implications of this study are manifold. Firstly, the BCRC model serves as an effective grading indicator and assistance for CSL learners. By accurately predicting the readability levels of texts, the model can guide learners in selecting appropriate reading materials that align with their language proficiency levels. This helps learners engage with texts that are neither too challenging nor too easy, fostering their reading comprehension and language acquisition abilities. Furthermore, this research has practical implications for practitioners in the field of international Chinese language and second and foreign language education. The BCRC model can aid educators in selecting more diverse and suitable reading materials for their students. It provides a reliable tool for educators to assess the difficulty of texts and make informed decisions about text selection, ensuring that learners are appropriately challenged and motivated in their language learning journey. Additionally, the model offers valuable guidance for textbook compilers in creating materials that cater to learners' reading abilities, promoting more effective and engaging learning experiences. Moreover, curriculum developers can benefit from this research as the BCRC model provides a useful reference for designing appropriate reading programs, ensuring a well-structured and comprehensive language curriculum for CSL learners. From a methodological perspective, this study offers insights into the effectiveness of using large language models for text grading. The BCRC model, built on the BERT architecture and trained on CSL texts, demonstrates the reliability and efficacy of using a large language model for text classification. The integration of linguistic features and the BERT model enhances the accuracy and validity of the BCRC model, providing a robust framework for assessing the readability of CSL texts.

There are a number of limitations to this study that need to be acknowledged and considered in future work. Firstly, the evaluation of the BCRC model could be further strengthened by using a larger and more varied dataset, encompassing different genres and topics of CSL texts. Additionally, this study mainly focused on linguistic features, and future research could explore the incorporation of other factors, such as cultural and contextual features, to gain a more comprehensive understanding of CSL text readability. Future work in this field should consider expanding the dataset to encompass a broader range of CSL texts and learner profiles, as well as incorporating reader-specific factors to enhance the accuracy and applicability of readability grading. Further research could also explore the combination of linguistic and

non-linguistic features to improve the predictive capability of the BCRC model. Additionally, incorporating user feedback and conducting longitudinal studies could contribute to refining and enhancing the readability grading of CSL texts.

REFERENCES

[1] S. Loewen, *Introduction to Instructed Second Language Acquisition*. New York, NY, USA: Routledge, 2015.
[2] D. Zhang and C.-H. Lin, Eds., *Chinese as a Second Language Assessment*. Singapore: Springer, 2017.
[3] C. R. Huang, J. S. Zhuo, and B. Meisterernst, Eds., *The Routledge Handbook of Chinese Applied Linguistics*. London, U.K.: Routledge, 2019.
[4] J. Jegerski, "Krashen and second language processing," *Foreign Lang. Ann.*, vol. 54, no. 2, pp. 318–323, Jul. 2021, doi: 10.1111/flan.12557.
[5] S. Krashen, *Second Language Acquisition and Second Language Learning*. Oxford, U.K.: Pergamon, 1981.
[6] D. K. Reed and S. Kershaw-Herrera, "An examination of text complexity as characterized by readability and cohesion," *J. Exp. Educ.*, vol. 84, no. 1, pp. 75–97, Jan. 2016, doi: 10.1080/00220973.2014.963214.
[7] Y. Sung, W. Lin, S. B. Dyson, K. Chang, and Y. Chen, "Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR," *Modern Lang. J.*, vol. 99, no. 2, pp. 371–391, Jun. 2015, doi: 10.1111/modl.12213.
[8] H. Eslami, "The effect of syntactic simplicity and complexity on the readability of the text," *J. Lang. Teaching Res.*, vol. 5, no. 5, pp. 1185–1191, Sep. 2014, doi: 10.4304/jltr.5.5.1185-1191.
[9] S. A. Crossley, J. L. Weston, S. T. M. Sullivan, and D. S. McNamara, "The development of writing proficiency as a function of grade level: A linguistic analysis," *Written Commun.*, vol. 28, no. 3, pp. 282–311, Jul. 2011, doi: 10.1177/0741088311410188.
[10] C. Ke, Ed., *The Routledge Handbook of Chinese Second Language Acquisition*. London, U.K.: Routledge, 2018.
[11] T. Yamasaki and K. Tokiwa, "A method of readability assessment for web documents using text features and HTML structures," *Electron. Commun. Jpn.*, vol. 97, no. 10, pp. 1–10, Oct. 2014, doi: 10.1002/ecj.11565.
[12] V. Solovyev, V. Ivanov, and M. Solnyshkina, "Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics," *J. Intell. Fuzzy Syst.*, vol. 34, no. 5, pp. 3049–3058, May 2018, doi: 10.3233/jifs-169489.
[13] C. Pires, A. Cavaco, and M. Vigário, "Towards the definition of linguistic metrics for evaluating text readability," *J. Quant. Linguistics*, vol. 24, no. 4, pp. 319–349, Oct. 2017, doi: 10.1080/09296174.2017.1311448.
[14] S. Phani, S. Lahiri, and A. Biswas, "Readability analysis of Bengali literary texts," *J. Quant. Linguistics*, vol. 26, no. 4, pp. 287–305, Oct. 2019, doi: 10.1080/09296174.2018.1499456.
[15] N. Nassiri, V. Cavalli-Sforza, and A. Lakhouaja, "Approaches, methods, and resources for assessing the readability of Arabic texts," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 4, pp. 1–30, Apr. 2023, doi: 10.1145/3571510.
[16] L. Jian, H. Xiang, and G. Le, "English text readability measurement based on convolutional neural network: A hybrid network model," *Comput. Intell. Neurosci.*, vol. 2022, Mar. 2022, Art. no. 6984586, doi: 10.1155/2022/6984586.
[17] X. Chen and D. Meurers, "Linking text readability and learner proficiency using linguistic complexity feature vector distance," *Comput. Assist. Lang. Learn.*, vol. 32, no. 4, pp. 418–447, May 2019, doi: 10.1080/09588221.2018.1527358.
[18] H.-C. Tseng, B. Chen, T.-H. Chang, and Y.-T. Sung, "Integrating LSA-based hierarchical conceptual space and machine learning methods for leveling the readability of domain-specific texts," *Natural Lang. Eng.*, vol. 25, no. 3, pp. 331–361, May 2019, doi: 10.1017/s1351324919000093.
[19] X. Sun and X. Huo, "Word-level and pinyin-level based Chinese short text classification," *IEEE Access*, vol. 10, pp. 125552–125563, 2022, doi: 10.1109/ACCESS.2022.3225659.
[20] A. Curiel, C. Gutiérrez-Soto, and J.-R. Rojano-Cáceres, "An online multi-source summarization algorithm for text readability in topic-based search," *Comput. Speech Lang.*, vol. 66, Mar. 2021, Art. no. 101143, doi: 10.1016/j.csl.2020.101143.
[21] S. Crossley, A. Heintz, J. S. Choi, J. Batchelor, M. Karimi, and A. Malatinszky, "A large-scaled corpus for assessing text readability," *Behav. Res. Methods*, vol. 55, no. 2, pp. 491–507, Mar. 2022, doi: 10.3758/s13428-022-01802-x.
[22] S. Nahatame, "Text readability and processing effort in second language reading: A computational and eye-tracking investigation," *Lang. Learn.*, vol. 71, no. 4, pp. 1004–1043, Dec. 2021, doi: 10.1111/lang.12455.
[23] G. R. Klare, "Assessing readability," *Reading Res. Quart.*, vol. 10, no. 1, p. 62, 1974.
[24] B. I. Koslin, S. Zeno, and S. Koslin, *The DRP: An Effective Measure in Reading*. New York, NY, USA: College Entrance Examination Board, 1987.
[25] A. J. Stenner, H. Burdick, E. E. Sanford, and D. S. Burdick, "How accurate are Lexile text measures?" *J. Appl. Meas.*, vol. 7, no. 3, pp. 307–322, 2006.
[26] D. S. McNamara, A. C. Graesser, P. M. McCarthy, and Z. Cai, *Automated Evaluation of Text and Discourse With Coh-Metrix*. New York, NY, USA: Cambridge Univ. Press, 2014.
[27] S. A. Crossley, D. B. Allen, and D. S. McNamara, "Text readability and intuitive simplification: A comparison of readability formulas," *Reading Foreign Lang.*, vol. 23, no. 1, pp. 84–101, 2011.
[28] R. G. Benjamin, "Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty," *Educ. Psychol. Rev.*, vol. 24, no. 1, pp. 63–88, Mar. 2012, doi: 10.1007/s10648-011-9181-8.
[29] Z. Yu, H. Li, and J. Feng, "Enhancing text classification with attention matrices based on BERT," *Expert Syst.*, vol. 41, no. 3, pp. 1–13, Mar. 2024, Art. no. e13512. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/exsy.13512, doi: 10.1111/exsy.13512
[30] Q. Mi, Y. Hao, L. Ou, and W. Ma, "Towards using visual, semantic and structural features to improve code readability classification," *J. Syst. Softw.*, vol. 193, Nov. 2022, Art. no. 111454, doi: 10.1016/j.jss.2022.111454.
[31] M. Bugueño and M. Mendoza, "Learning to combine classifiers outputs with the transformer for text classification," *Intell. Data Anal.*, vol. 24, pp. 15–41, Dec. 2020, doi: 10.3233/ida-200007.
[32] Y. Cui, "CTAP for Chinese: A linguistic complexity feature automatic calculation platform," in *Proc. Lang. Resour. Eval. Conf.*, 2022, pp. 5525–5538. [Online]. Available: https://aclanthology.org/2022.lrec-1.592
[33] Y. Zhang, J. Song, W. Peng, D. Guo, and T. Song, "A machine learning classification algorithm for vocabulary grading in Chinese language teaching," *Tehnicki Vjesnik, Tech. Gazette*, vol. 28, no. 3, pp. 845–855, 2021, doi: 10.17559/TV-20210128043310.
[34] X. Sun, Z. Liu, and X. Huo, "Six-granularity based Chinese short text classification," *IEEE Access*, vol. 11, pp. 35841–35852, 2023, doi: 10.1109/ACCESS.2023.3265712.
[35] X. Liu, S. Wang, S. Lu, Z. Yin, X. Li, L. Yin, J. Tian, and W. Zheng, "Adapting feature selection algorithms for the classification of Chinese texts," *Systems*, vol. 11, no. 9, p. 483, Sep. 2023, doi: 10.3390/systems11090483.
[36] M. Liu, Y. Li, Y. Su, and H. Li, "Text complexity of Chinese elementary school textbooks: Analysis of text linguistic features using machine learning algorithms," *Sci. Stud. Reading*, vol. 28, no. 3, pp. 235–255, May 2024, doi: 10.1080/10888438.2023.2244620.
[37] Y.-T. Sung, J.-L. Chen, J.-H. Cha, H.-C. Tseng, T.-H. Chang, and K.-E. Chang, "Constructing and validating readability models: The method of integrating multilevel linguistic features with machine learning," *Behav. Res. Methods*, vol. 47, no. 2, pp. 340–354, Jun. 2015, doi: 10.3758/s13428-014-0459-x.
[38] D. Luo, J. Gong, and Y. Li, "The effect of reading comprehension questions on linguistic predictors of readability," *Educ. Stud.*, vol. 48, no. 5, pp. 659–675, Sep. 2022, doi: 10.1080/03055698.2020.1798741.
[39] Ministry of Education PRC. *Chinese Proficiency Grading Standards for International Chinese Language Education*. Accessed: Mar. 31, 2021. [Online]. Available: http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202103/t20210329_523304.html
[40] L. Ortega, "Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing," *Appl. Linguistics*, vol. 24, no. 4, pp. 492–518, Dec. 2003, doi: 10.1093/applin/24.4.492.

[41] M. Kim and S. A. Crossley, "Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing," *Assessing Writing*, vol. 37, pp. 39–56, Jul. 2018, doi: 10.1016/j.asw.2018.03.002.
[42] J. Read, *Assessing Vocabulary*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[43] X. Lu, "The relationship of lexical richness to the quality of ESL learners' oral narratives," *Mod. Lang. J.*, vol. 96, no. 2, pp. 190–208, Jun. 2012, doi: 10.1111/j.1540-4781.2011.01232.x.
[44] K. Wolfe-Quintero, S. Inagaki, and H.-Y. Kim, *Second Language Development in Writing: Measures of Fluency, Accuracy, & Complexity*. Honolulu, HI, USA: Second Language Teaching & Curriculum Center, 1998.
[45] J. Ure, "Lexical density: A computational technique and some findings," in *Talking About Text*, M. Coultard, Ed. Birmingham, U.K.: Univ. of Birmingham, 1971, pp. 27–48.
[46] L. Ortega, "Syntactic complexity in L2 writing: Progress and expansion," *J. 2nd Lang. Writing*, vol. 29, pp. 82–94, Sep. 2015, doi: 10.1016/j.jslw.2015.06.008.
[47] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, Baltimore, MD, USA, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010
[48] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[49] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html
[50] F. Pedregosa, S. Varoquaux, A. Gramfort, V. Michel, and B. Thirion, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Dec. 2011.
[51] F. Kuiken and I. Vedder, Eds., *Syntactic Complexity, Lexical Variation and Accuracy as a Function of Task Complexity and Proficiency Level in L2 Writing and Speaking*. Amsterdam, The Netherlands: John Benjamins, 2012.

CHAO ZHANG received the Ph.D. degree in Chinese language and literature, in 2022, with a focus on applied linguistics research.

He is currently an Associate Professor with the College of Foreign Languages, Qufu Normal University, China, where he is also a Master's Supervisor and the Deputy Director of the Center for Hong Kong and Macau Education Research and the Digital Humanities Research Center. He has published more than ten peer-reviewed journal articles in prestigious domestic and international SCI/SSCI-indexed journals, such as *Language Teaching Research*, *Frontiers in Psychology*, and *Complexity*, where he also serves as a Peer-Reviewer. Additionally, he has participated in several national-level research projects. His primary research interests include second language writing, vocabulary learning, and readability leveling.

Dr. Zhang is an active member of several professional societies and has received recognition for his academic contributions, including serving as a peer-reviewer for international SCI/SSCI journals. His research findings have been presented at various national and international conferences, contributing to the advancement of knowledge in the field of second language acquisition and pedagogy.