Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 2715408, 11 pages
https://doi.org/10.1155/2022/2715408

*Research Article*

# Constructing a Data-Driven Model of English Language Teaching with a Multidimensional Corpus

**Dongyan Chen** [1,2]

[1]*College of International Studies, Beibu Gulf University, Guangxi 535015, China*
[2]*Academy of Language Studies, University of Technology MARA, Negeri Selangor 40450, Malaysia*

Correspondence should be addressed to Dongyan Chen; chendongyan1120@163.com

Received 21 February 2022; Accepted 28 March 2022; Published 28 June 2022

Academic Editor: Gengxin Sun



In this study, a multidimensional corpus English teaching model is constructed using the data-driven model. This study uses a data-driven collection of massive amounts of data to generate a multidimensional corpus. The data-driven generation of a multidimensional corpus to build a teaching model is studied, and the principle of data driven, the computational process, and the characteristics of the corpus are analyzed. Due to the deficiency of data-driven modeling without correlating process variables with quality variables, this study adopts an artificial intelligence algorithm and analyzes the basic principle, computational process, and advantages and disadvantages of the method. The model is simulated and verified for multidimensional corpus and English teaching. To address the shortcomings of the AI algorithm, which has a complex computation process and no orthogonal decomposition of the data space, the autoregressive latent structure projection algorithm is designed by integrating the autoregressive idea with the artificial intelligence (AI) algorithm. This algorithm can orthogonally decompose the sample data space and simplify the modeling process. Finally, the algorithm is validated by simulation. To verify the results of the teaching model, the fuzzy C-means clustering algorithm is combined with the autoregressive latent structure projection algorithm in this study. The sample data used in the modeling are divided into categories, and the affiliation function is calculated for each category. The affiliation function is used to calculate the affiliation of the online calculation results for each category, and the final evaluation results are obtained based on the fuzzy comprehensive evaluation method. Finally, taking junior students as an example, the simulation is carried out to verify the effectiveness of the English teaching model. The research results show that the corpus-based English flipped classroom teaching model improves English teaching methods, enhances students' English proficiency and independent learning ability, and provides a practical basis for English teaching model exploration.

## 1. Introduction

Technology continues to penetrate the field of education, and students interact with various platforms, generating a large amount of learning behavior and achievement data, which have significant educational value when accumulated over time [1]. When teachers teach and research, they should actively explore and use student data to diagnose student problems and improve teaching methods so that teachers can transform their teaching and research and teaching from empirical and process oriented to scientific and personalized. The application of education data in education can promote the development of education informatization and education modernization. At present, data analysis platforms are gradually being built in primary and secondary schools, and educational data are being accumulated. Teachers can use student data to gain a deeper understanding of students' learning needs, gain insight into the path of learners' learning behavior to improve teaching, enhance teaching effectiveness, and promote teachers' professional development [2]. Data-driven teaching has been the frontier of international education information development, and data-driven teaching has four characteristics: scientific, precise, intelligent, and personalized. Data-driven teaching and research are the links before data-driven teaching is carried out, teaching and research are the foundation of teaching, and scientific teaching and research are conducive to improving the teaching effect. Although

some schools have not built data platforms, with the increasing awareness of data use, teachers should have an in-depth understanding of the data-driven teaching and research process in advance from the principle of understanding the process of data generation, acquisition, processing, and analysis. At present, the theoretical research on multidimensional corpus English teaching and research in China are insufficient [3]. There is not enough theoretical research on multidimensional corpus ELT teaching and research in China to guide teachers in corpus teaching and research practice. The purpose of this study is to study the data-driven corpus teaching and research process, to develop a corpus English teaching and research program, and to guide teachers step by step to utilize the corpus of student data so that student data can help teachers in their teaching decisions.

As the national demand for high-end foreign language talents continues to increase, foreign language education is facing higher and higher requirements for talent cultivation, and the cultivation of international composite talents with "one specialization and multiple abilities" and "one proficiency and multiple skills" has become one of the important directions of foreign language education reform in universities [4]. In the process of cultivating complex talents, academic English teaching plays an important role, because academic English highlights the instrumental characteristics of English, which can meet the practical needs of students' professional study and cultivate students' ability to use English for work and scientific research. In recent years, the process of national education informatization has continued to advance, and information technology has revolutionized higher education, especially foreign language education, triggering deep changes in foreign language education philosophy, teaching organization, and teaching methods [5]. The Guide to Teaching English at University requires teachers to build and use microcourses and catechisms, use online high-quality educational resources to transform and expand teaching contents, and implement hybrid teaching modes such as flipped classes based on classroom and online courses [6]. Thus, this study tries to construct a corpus-based English flipped classroom teaching model, empirically test its teaching effect, and provide a practical reference for English teaching model exploration.

## 2. Related Works

Data-driven teaching and research abroad originated at the beginning of the 21st century and were earlier called professional learning communities (PLCs) and later called data teams (DTs). WestEd is a nonprofit research, development, and service organization dedicated to improving the education of children, youth, and adults. Led by Ellen Mandinach, senior research scientist and director of decision data, WestEd researchers believe that teachers should not be taught data literacy skills in isolation; they believe that data literacy skills should be taught in conjunction with data use processes [7]. Jamal et al. also conducted an empirical study of the impact of data team implementation on student achievement by implementing the Harvard Data Wisdom Improvement Process at Leasure Elementary School, led by the school's leadership, to train the school's teachers on how to organize collaborative work, lead teachers to dig deeper into student data to identify instructional problems, and then create instructional solutions for students and teachers based on the team's findings [8]. After implementation, Leasure Elementary's test scores reached their highest level since 2009, bridging the gap between the test scores of special education students and the general student population. In the direction of teacher data literacy research, various experts and scholars offered their insights on data literacy. Zhang and Han believe that teacher data literacy consists of three major components: data awareness, data competence, and data ethics [9]. Di Gangi used the ACTS academic quality evaluation report form to develop new teaching strategies by first reading the data, identifying problems, focusing on them, analyzing the causes, addressing them, and exploring the teaching in three steps so that the average score of this class changed from below the regional average to above the average, which verified the effectiveness of data-driven teaching research and highlighted the application value of student achievement data This demonstrates the effectiveness of data-driven teaching and research and highlights the value of student achievement data [10]. According to the law of large numbers, when the number of training samples tends to be infinite, the empirical risk approaches the expected risk, and the prediction can be accurate for new samples.

European linguists and educators proposed the use of corpora as an aid in the teaching and learning process of foreign languages. The corpus is a very important branch of corpus linguistics that can be used as a teaching aid in foreign language teaching and is considered an effective teaching method that encompasses two main aspects: one is to directly teach corpus knowledge, using the corpus as a means and resource for language teaching; the other is to indirectly use it as a tool for lexicography, grammar reference, grammar teaching, and other multimedia courseware and as a corpus and computer-based language learning software and testing tool. Tsai, a linguist, made an important contribution to corpus applications, arguing that vocabulary teaching is the primary task of foreign language teaching [11]. At the beginning of the 21st century, Tsai again advocated the use of corpus-based chunking in foreign language teaching. He suggests that learners independently learn through authentic corpora so that students can understand the meaning and usage of vocabulary. This is a typical example of the application of the corpus-based block teaching method in language teaching, emphasizing the authenticity of the corpus and the scientific nature of the computer as a supplementary teaching tool in language teaching [12]. Piotrkowicz et al. analyzed the use of chunks in Chinese English learners' language output using the "corpus of Chinese English learners' spoken language" and showed the quantity and quality of chunks used by Chinese students in both languages' organization and content selection that are not satisfactory [13]. Hooshyar et al. argue that the corpus is more applicable to higher education and that the richer and faster updating of the corpus in the case

of postgraduate academic learning proves that the corpus can better help students in international learning and communication [14].

## 3. Construction of a Data-Driven Multidimensional Corpus-Based ELT Model

*3.1. Data-Driven Model Design.* As we all know, the data-driven algorithm is different from a model-driven algorithm; in that it no longer needs to build a physical model for a specific problem but uses the data generated by the problem for various tasks such as monitoring, evaluation, and control. Artificial intelligence (AI), as the mainstream algorithm of data-driven technology, has made breakthroughs in four aspects, such as algorithm, data, computing power, and framework, in the past two decades, and thus has been widely used in various fields [15]. Artificial intelligence algorithms can be mainly classified into three categories: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is the most used type of AI algorithm, usually using a training set consisting of $n$ "input-output" pairs $\{(x_i, y_i)_{i=1}^{n}\}$ to continuously "learn" the mapping relationship between input and output $y = f(x)$. Supervised learning is essentially the process of "fitting" the mapping relationship, where the optimization parameters in the mapping are continuously adjusted by an optimization algorithm to reduce the "loss" value of the training samples, also known as empirical risk, i.e.,

$$\begin{aligned} Loss(f) &\le f(x)\sum e^{x}, \\ y_i &= t_j x_i + \frac{x_j}{t_{i+1} + t_j}. \end{aligned} \tag{1}$$

According to the law of large numbers, when the training samples tend to infinity, the empirical risk tends to be closer to the expected risk, and thus, the prediction can be accurate for new samples. Common supervised learning algorithms include shallow common algorithms such as decision trees (DTs), support-vector machines (SVMs), and neural networks (NNs), and convolutional neural networks (CNN), recurrent neural network (RNN), and other deep neural networks. In this study, deep neural networks are mainly used, but ordinary neural networks are also used for some applications where the mapping relationship is relatively simple.

To build an AI model that meets the practical application requirements using the above algorithms, we need to realize the tedious and underlying code development such as data storage, model building, optimization training, and hardware acceleration. It is impractical to go through such a tedious development process for each model building. Therefore, open-source or commercial software frameworks have been developed to address this problem, which can be used by developers in related fields to accelerate the process of model building, data access, and training inference. Software frameworks include both hardware acceleration of software frameworks developed by hardware vendors for deep learning and software frameworks that focus on model building and fast training for developers.

The principal component analysis is often applied to multimetric performance evaluation, which can solve the problem of the high complexity of analysis and evaluation caused by the excessive number of metrics in the process of multimetric evaluation. Due to the strong correlation of indicators in the original data, the information reflected by indicators will overlap. The principal component analysis (PCA) algorithm uses as few independent new indicators as possible for the original sample data and reflects the process information carried by the original sample data as much as possible, through which the correlation between the indicators of the original sample data is eliminated and the dimensionality of the original sample data is reduced. The principle of the algorithm is shown in

$$\begin{aligned} X &= [x_1, x_2, \ldots, x_n] \\ &= \begin{Bmatrix} x_1 & \ldots & x_n \\ x_{11} & \ldots & x_{1n} \\ x_{n1} & \ldots & x_{nn} \end{Bmatrix}, \end{aligned} \tag{2}$$

$$\mathrm{STD}_{ij} = S_i + \frac{X_j}{x_{ij}}. \tag{3}$$

The current breakthrough in data-driven technology is due to the improvement of the "source of intelligence" algorithm, rather than the accumulation of massive data and the improvement of computing power, which are the most important factors for breakthroughs in the field of artificial intelligence in the past decade. Data are the core of data-driven algorithms, and the scale and quality of the data directly affect the accuracy and generalization ability of the trained models. To carry out the task of transient stability assessment using massive data samples, researchers have tried various data-driven algorithms in recent years to continuously improve the accuracy and computational efficiency of transient stability assessment. In general, there are two main types of ideas for constructing stability assessment models with the help of data-driven algorithms: one is to construct stability boundaries with the help of massive data, based on which the stability conclusions are drawn by judging the relative position of the current operating state and the boundaries. It is true that some algorithms skip the step of boundary construction and directly draw stability conclusions based on some existing data samples in the relative neighborhood. Another class of ideas is to construct mapping relations from system measurement information or system operation and disturbance characteristics to stability conclusions with the help of massive data samples. Although the constructed mapping relations are like stability boundaries in mathematical essence, they have significant differences in the way of thinking and are, therefore, considered as another class.

To enhance the feature extraction performance of the feature extractor, its order and parameters must be carefully designed with the help of convolutional, activation, and pooling layers. In general, the deeper the neural network

structure, the higher the accuracy of the model on more complex tasks, but it also leads to longer training time, poorer convergence characteristics of optimization, and more severe overfitting. Therefore, the essence of deep neural network model design is a trade-off between model complexity and accuracy. In common classical models, convolutional and activation layers are usually combined to extract features from the input. It is obviously unrealistic to go through such a tedious development process for each model building. To exactly determine how many convolution and activation layers are needed, these two layers can be added to the model until the accuracy of the model no longer significantly improves. At the same time, pooling layers can be added to the model to reduce the training parameters without significantly sacrificing the accuracy of the model. Based on this principle, the convolutional neural network feature extractor section shown in Figure 1 is designed and used in the example analysis section of this study: it contains five convolutional layers, five activation layers, and three pooling layers. The critical line's transfer power must not exceed the available transfer capability (ATC), and the feature extractor is used to discover more implicit "rules" to distinguish the stability of each sample to be evaluated.

After extracting the input feature information, the network structure of the fully connected layer is constructed to establish the mapping between the features extracted from the training samples and their stability findings to predict the stability findings of the new samples. Most deep learning models, including AlexNet and VGG, use a network structure with 3 layers of fully connected layers. Considering the complex and high-dimensional nonlinear nature of the transient stability problem, the number of layers is set to 3 here. In addition, considering that the fully connected layer is prone to overfitting during training, the dropout technique is used here to enhance the robustness of the network by forcing some hidden neurons to zero.

*3.2. Building a Multidimensional Corpus ELT Model.* A corpus (plural corpora), as its name implies, is a storehouse of linguistic materials, a database of written and spoken language stored in a computer for research purposes. The distinguishing feature of a corpus is that the language materials are real materials in actual use, covering a wide range of fields such as literature, business, and educational translation, and the number of words covered is in the hundreds of millions. The corpus is preferred by scholars in various fields because of the observability and verifiability of the data it provides, its shareability, and its ease of retrieval [16]. Corpora use random sampling methods to collect naturally occurring continuous language according to certain linguistic rules, through texts or language fragments to build a large electronic textbase with a certain capacity. Currently, corpus linguistics is widely used in foreign language research. The research in foreign language teaching is divided into two types. On the one hand, there are applied studies that use the materials in the corpus as research texts. This type of research mainly uses learner corpora, such as the Chinese Learner English Corpus (CLEC) and the International Corpus of Learner English (ICLE), to conduct a comprehensive study of various lexical or grammatical error features in the writing or speaking of English learners at home and abroad. To determine how many convolution and activation layers are needed, you can keep adding these two layers to the model until the accuracy of the model no longer significantly improves. At the same time, pooling layers can be added to this model to reduce training parameters without significantly sacrificing model accuracy.

Microlearning is a product of the deep integration of information technology and education teaching and has become an important teaching resource. With the characteristics of "prominent theme, short and concise, interesting, and wide application," microlessons help share high-quality teaching resources and make learning possible anytime and anywhere. To meet the fragmented learning needs of students, we have developed a series of microlessons on academic English vocabulary, grammar, reading, translation, writing, and listening, guided by the principles of integration of academic and interest, unification of thematic and application, and coordination of focus and relevance. Based on the corpus-based academic English teaching platform, we have built a flipped English classroom teaching model, as shown in Figure 2, which aims to support the corpus-based teaching platform and combines the advantages of traditional classrooms to interconnect online interactive teaching and offline interactive teaching to form a flipped classroom organism, so that the time and space of English teaching and learning can be infinitely extended and the purpose of improving students' English ability and academic literacy can be achieved.

The basic idea of partial least squares is to decompose the process variable data space into two subspaces according to the magnitude of correlation with the quality variable $X$, i.e., the subspace X' containing the correlation between the process variable $X$ and Y, and the residual matrix X0, which is uncorrelated with the quality variable Y. However, the nonlinear iterative partial least squares (NIPALS) algorithm used in this algorithm has difficulty in ensuring that $X'$ and X0 are mutually orthogonal, and the algorithm loops once to obtain a score vector, leading to high computational complexity. YIN et al. proposed an autoregressive projection to latent structure by combining autoregressive ideas with partial least squares. The algorithm uses historical data, establishes the corresponding regression coefficient matrix, and performs orthogonal decomposition of the sample space of historical data of process variables based on the principle of the magnitude of the correlation between process variables and quality variables. The algorithm can solve the problem of the high complexity of standard partial least squares operation. With the complete decomposition of the quality variable data space Y, the matrix can specifically reflect the correlation between the process variable $X$ and the quality variable Y, which is defined as the regression coefficient matrix according to the form of the regression algorithm. e' denotes the residual space of the quality variable data space Y. With a complete decomposition of the quality variable data space, $E'$ should be independent of the process variable data space $X$.

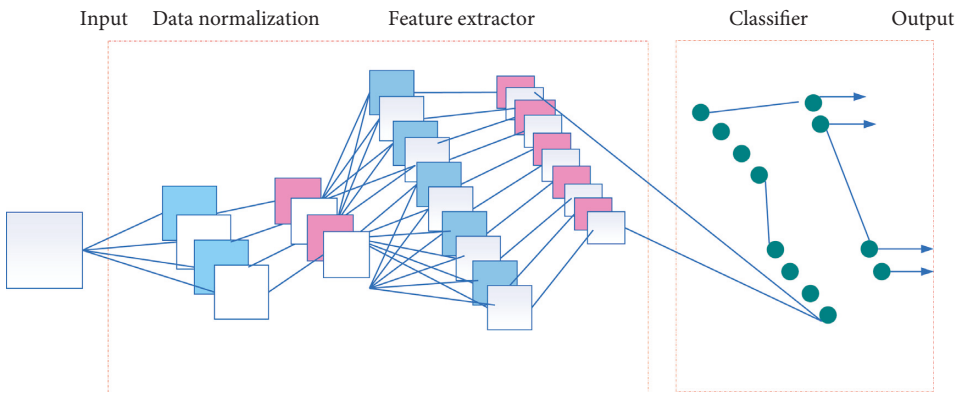


Figure 1: Structure of the feasible convolutional neural network.

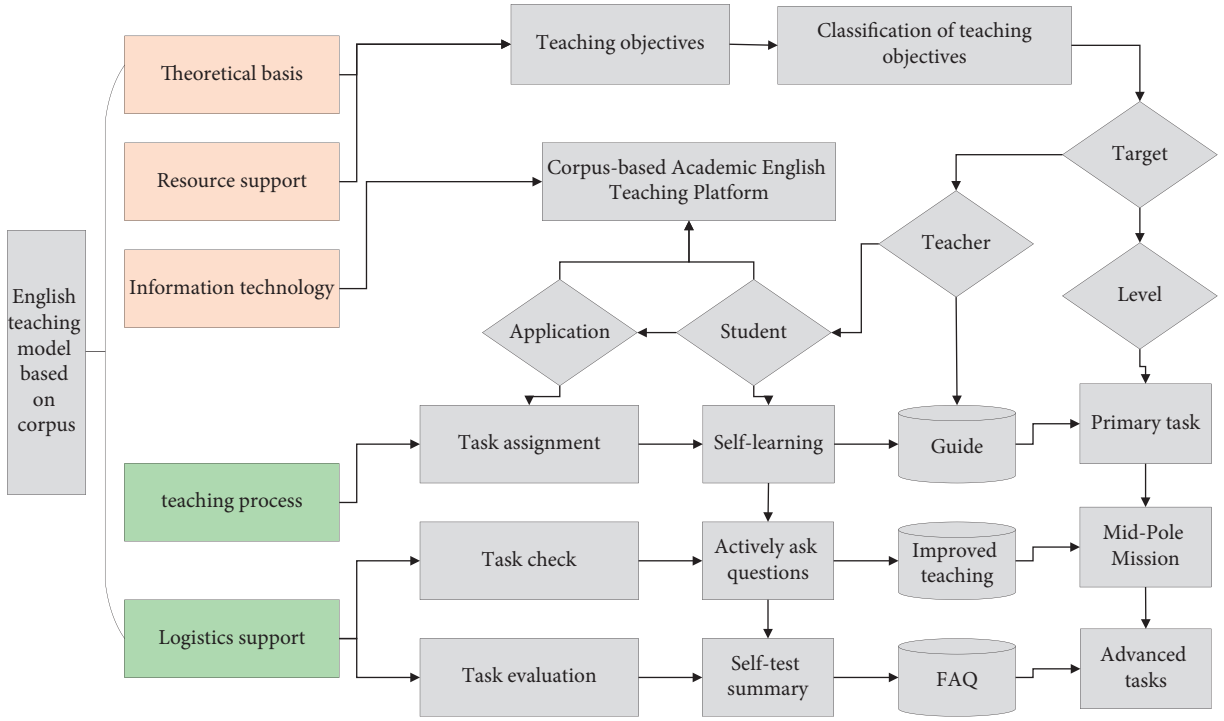


Figure 2: English teaching model based on the corpus.

$$E\left(e_y, x\right) = cov\left(e_y x^T\right) \le 1,$$
$$X = X' + \hat{X} \quad (4)$$
$$= P_M P_t^M - X' P' P_M^t.$$

From Figure 3, it is easy to find that when the number of latent variables $p > 2$, the modeling process is more stable because the autoregressive latent structure projection (AR-PLS) algorithm does not use the nonlinear iterative partial least squares (NIPALS) method used in regression modeling. In addition, the standard partial least squares method requires a given number of latent variables for modeling, and the determination of the number of latent variables has a significant impact on the process monitoring. The calculation process of the algorithm is simple, easy to understand, and has high efficiency in calculating high-dimensional sample data. The fuzzy C-means algorithm converts the traditional fuzzy clustering method into an optimization problem with a constraint function. There is no theoretical method to determine the number of latent variables, and more practical methods such as cross-validation methods are used to determine the number of latent variables, which brings uncertainty to the established industrial process monitoring models. In contrast, the autoregressive latent structure projection (AR-PLS) algorithm no longer requires the number of latent variables to be set, which has certain advantages, as shown in Figure 3.

To clarify the research progress of data-driven teaching and research models, foreign teaching and research models were sorted out. At present, foreign research on data-driven teaching and research models is more mature, and foreign countries call data-driven teaching and research models as collaborative data team procedure, datawise process, etc.

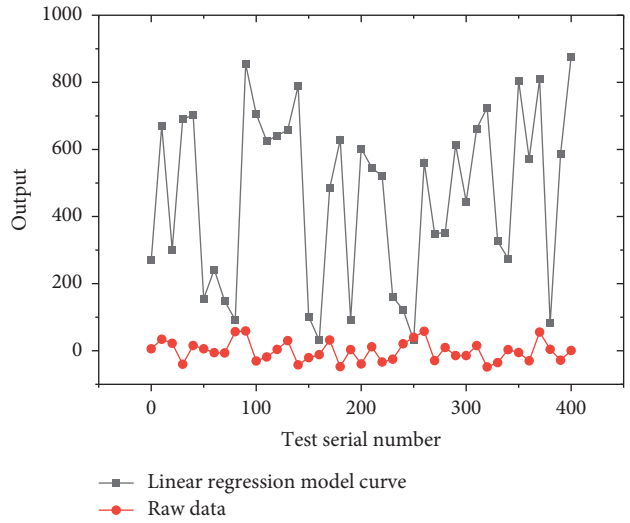


Figure 3: Autoregressive latent structure projection model performance test results.

Kim Schildkamp et al. are dedicated to studying how schools use student data and how to organize and build data teams so that they can better serve teaching and learning. Mandinach developed a complete approach to data use, data literacy for teachers (DFLT), and a conceptual framework for teachers. As shown in Figure 4, content knowledge, curriculum knowledge, learners' knowledge and learner characteristics, awareness of educational purposes and values, general pedagogical knowledge, pedagogical content knowledge, and educational background knowledge are used as inputs for the data use of the teaching process.

In the traditional machine learning approach, model modeling requires the provision of three datasets: training, validation, and testing. These are applied to model fitting, model hyperparameter tuning, and assessing model generalization capabilities, respectively. Of course, it is the best choice if the required experimental information can be collected in the real physical network, but data collection in the real network faces problems such as the impact of real physical network information collection on the performance of the present network, the high cost of real physical network information collection, and the topology of the real physical network topology in the case of guaranteed homogeneity with a single topology [6]. However, the nonlinear iterative partial least squares (NIPALS) algorithm used in this algorithm is difficult to ensure that X′ and X0 are mutually orthogonal, and to obtain a score vector, the algorithm loops once, resulting in high computational complexity. Therefore, the network simulation service hopes to construct the network simulation dataset required for modeling the network delay performance algorithm in the network delay inference service through a discrete event-driven network simulator.

## 4. Results Analysis

*4.1. Data-Driven Model Results.* The network simulation dataset should be able to meet the modeling requirements of the network delay performance algorithm in the network delay inference service. On the other hand, the network managers in the inferred system management platform need to follow the standards defined in the network simulation service for a specific network scenario, and only through the inferred system management platform can the network managers complete the use of network delay performance prediction and historical information query for a specific network scenario [17]. The network simulation dataset built by the network simulation service runs through the entire development and uses the process of the data-driven network QoS inference system, which puts higher requirements on the reliability of the network simulation dataset.

Cluster analysis is an important research element in the field of data mining, and cluster analysis algorithms are widely used in many fields of daily life. Clustering analysis can explore the structural features inside the data, classify and label the generated data, and then uncover the potential and unknown information inside the data. Cluster analysis is a classical unsupervised classification method that uses mathematical methods to automatically classify a sample set of data without giving classification principles in advance. The fuzzy C-means cluster analysis algorithm can solve the requirement of fuzziness, which is difficult to solve by hard cluster analysis methods and is a coarse division of sample data. The algorithm has a simple and easy-to-understand computational process and has high efficiency in computing high-dimensional sample data. The fuzzy C-means algorithm converts the traditional fuzzy clustering method into solving optimization problems with constraint functions. As shown in Figure 5, the use of the higher-order data-driven arbitrary polynomial chaos expansion method implies the use of the higher-order polynomial basis functions in polynomial approximation, and more polynomial basis functions will bring closer approximation to the simulation results.

The performance evaluation based on the autoregressive latent structure projection algorithm proposed in this study first uses the production process prediction model established by the autoregressive latent structure projection algorithm, then uses the fuzzy C-means cluster analysis algorithm to calculate the affiliation of the output variable data in the modeled data for each performance level, and obtains the affiliation function of each variable for each performance level [11]. In the performance evaluation of online data, due to the lag in the production process, the output variables are first predicted using the prediction model and the input variable data, and then, the affiliation function is used to calculate the affiliation of each output variable predicted value for each performance level, and finally, the performance level belonging to that moment is obtained using the fuzzy operator to provide a reference for the field operators. It is not difficult to find from Figure 3 that when the number of latent variables $p > 2$, since the autoregressive latent structure projection (AR-PLS) algorithm does not use the partial least squares method, the nonlinear iterative partial least squares (NIPALS) method used in regression modeling makes the modeling process more stable.

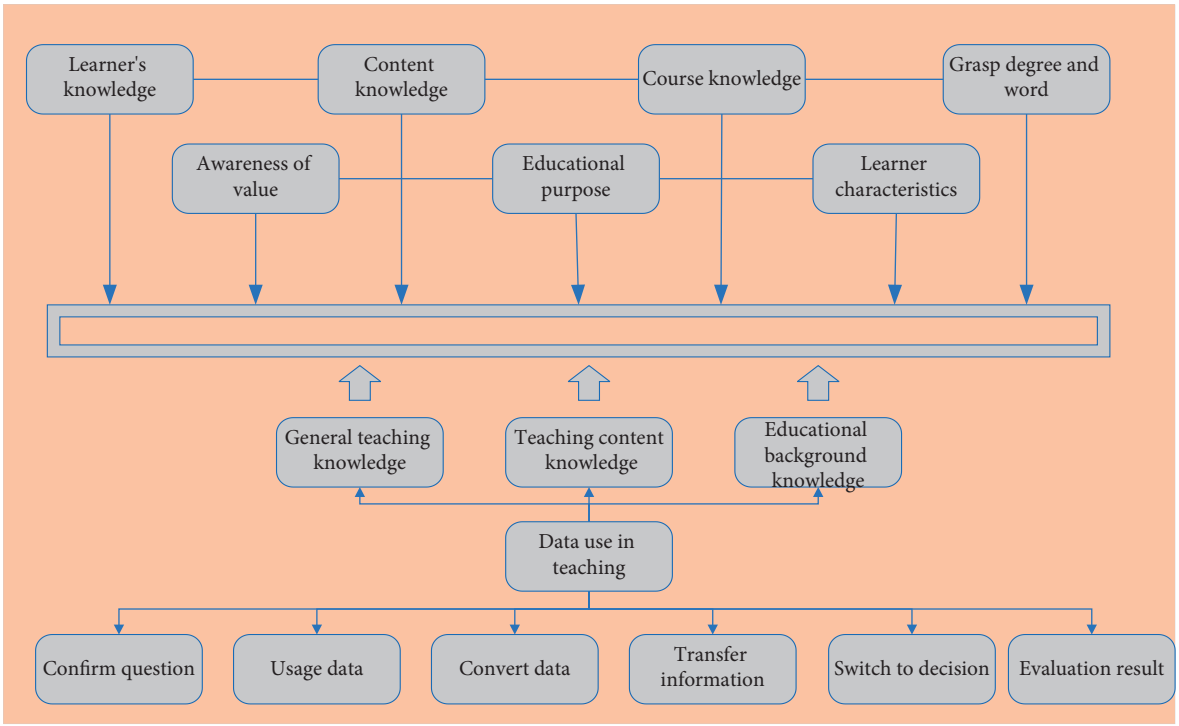


Figure 4: Data literacy of teachers' conceptual framework.

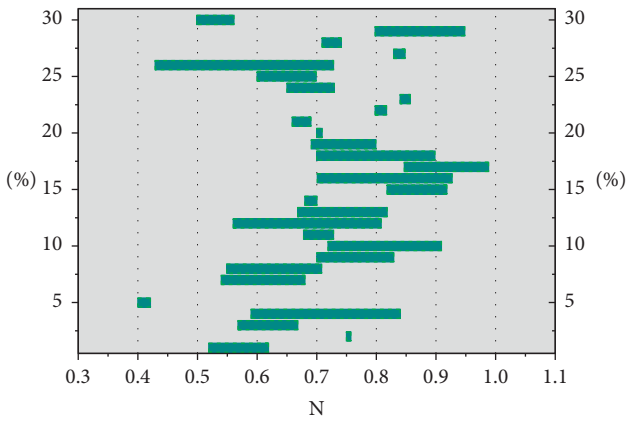


Figure 5: Eigenvalue method to analyze the number of clusters.

The ecological foreign language teaching model in the computer network environment is a theoretical teaching procedure that uses the theory of foreign language teaching and educational ecology as the common support theory, the optimal combination of all teaching elements as the construction base, the computer network technology as the development intermediary, the use of different teaching strategies to maximize the presentation of teaching content, to achieve the set teaching goals, the structural framework of various teaching activities, and the collection of teaching methods. It is the direction of rational construction and optimization of modern foreign language teaching mode. This model advocates open teaching information selection according to the teaching needs and mostly carries out task-based teaching activities in the form of collaboration and mutual assistance between teachers and students, which helps to cultivate students' language communication and comprehensive application skills. Because of the current development of the integration of computer network technology and foreign language courses, we must comprehensively compare and analyze the ideal one under the premise of considering various factors such as hardware and software conditions of computer network technology, teaching objectives of foreign language courses, teaching staff's willingness to choose information and network teaching literacy, students' learning motivation, level and network application ability, and auxiliary background management mechanism of network teaching. To build a foreign language teaching model that meets its development conditions has clear development goals, and wide development space, a reasonable series of planning and design can be carried out. The performance test results of the data-driven model are shown in Figure 6.

In this chapter, a framework of data-driven transient stability boundary generation and an online stability evaluation algorithm are proposed. To this end, a critical transient stability sample sampling and resampling mechanism is first proposed to accelerate the generation of sufficient critical transient samples in the high information entropy region near the stability boundary to provide a data basis for transient stability boundary generation. In addition, a critical operation and perturbation scenario screening mechanism is developed to further reduce the search space of the system, which provides a feasible solution to the challenges of "dimensional disaster" and "combinatorial explosion" faced by the stable boundary construction problem. Overall, the algorithm not only significantly

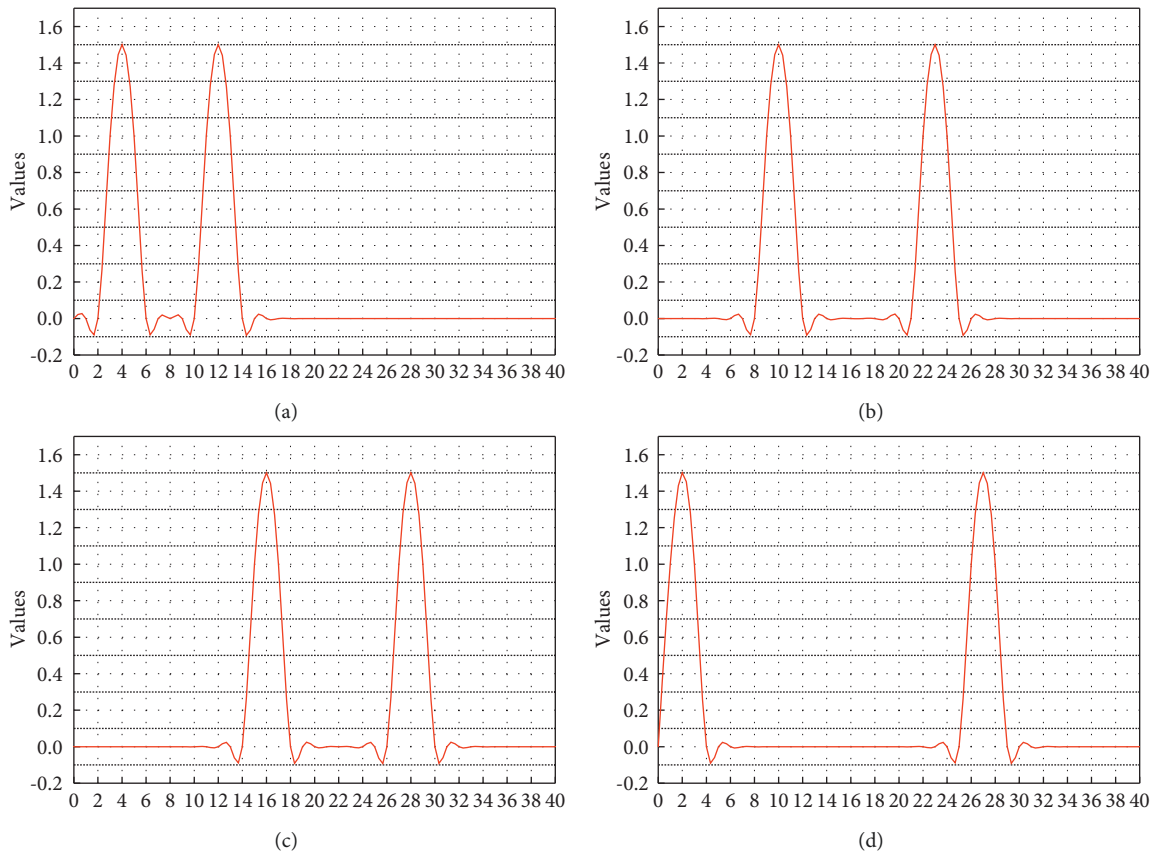


Figure 6: Data-driven model performance test results.

improves the efficiency of key transient sample generation and accelerates the speed of transient stable boundary generation but also significantly reduces the computational burden with the designed running point tracking and periodic boundary update mechanism, which makes it possible to update the state of the multidimensional corpus in real time based on data driven. In addition, the standard partial least squares method requires a given number of latent variables in modeling, and the determination of the number of latent variables has a great impact on process monitoring.

*4.2. Simulation Experiments of Multidimensional Corpus English Teaching Model.* The new computer network teaching environment requires teachers to change their teaching concepts, adhere to the "student-centered" teaching concept, act as an analyst of students' learning needs, a guide of learning direction, and a supervisor of overall activities, and assist students to realize the transformation of the central subject position of classroom teaching activities and the improvement of independent learning ability on the internet [18]. The students' learning needs are analyzed, learning directions are guided, and overall activities are supervised. However, research statistics show that only 47.6% and 44.3% of teachers and students, on average, believe that teachers can act as directional guides in multimedia classrooms and students' online learning, respectively, which shows that more than half of teachers have almost neglected their leading role. If they cannot design classroom activities, they cannot provide students with opportunities for active thinking and feedback, and they cannot effectively motivate students to learn. At the same time, as many as 43.6% and 47.3% of students are evaluated as marginalized in multimedia classroom teaching activities and passive recipients of knowledge in online independent learning, resulting in a serious lack of their role as learning subjects. Their personal learning needs are not clear, their ability to make independent decisions is poor, and the learning process is only blindly and passively accepted. There are even 34.7% of students who are evaluated as internet information losers, and their learning effect is even worse.

For this study, we selected English III students from the university of $X$ to implement this data-driven teaching and research project. University of $X$ is a second-level undergraduate institution with a strong overall student body and a strong faculty and has sufficient capacity to complete the experiment. As shown in Table 1, five teachers participated in this school, one as the subject team leader, one information system management teacher, and two subject teachers. The information system management teacher did not participate in the entire data-driven teaching and learning activity; the teacher participated in the data-driven teaching and learning in stage 3 by providing support for the teaching and learning data survey. Student data support was needed when conducting student problem queries. Querying student data was not easy, the school did not have all student

Table 1: Composition of teachers in the experimental group.

| Participant | Function | Teaching age | Subject | Requirements or not |
|---|---|---|---|---|
| Director Wang | Subject group leader | 10 | English | Yes |
| Teacher Ma | Teacher | 6 | English | Yes |
| Ms. Zhao | Teacher | 6 | Science | Yes |
| Mr. Xu | Teacher | 7 | English | Yes |
| Instructor Sun | Data coach | — | Computer science | Yes |

data recorded in an electronic system, and access to the system administrator revealed that the school had a data system, but only stored midterm and final grades and text-based teacher evaluations of students' each semester. This study led teachers to record weekly paper test scores in electronic form so that teachers could use the data for visual analysis. If the school had more comprehensive data, it would support teachers in making more accurate decisions, and the lack of easy access to data seriously affects the effectiveness of data-driven teaching and research.

First, the data coach explained the types of teaching data for teachers, and then, the data coach instructed the three teachers and the data system management teacher to recall and query the data stored in this grade in our school to record the data of this data-driven teaching and research class. This data record form records the data of this grade level in seven aspects: data number, data name, data generation time, data form, data storage location, public object, and data use. The four main types of data related to teaching and learning in grade 4 are unit test scores, student classroom performance ratings, student grouping data, formal routine assessments, and formal classroom assessments. From the unit test scores, we can diagnose the weak points of students' knowledge, and by correlating student grouping data, student classroom performance scores, and unit test scores, we can explore the degree of correlation between students' usual performance and grades. In short, these data are the basis for our data-driven teaching and research.

Before the experiment began, the author administered a controlled output vocabulary test in both the experimental and control classes, and after the test, the test scores were entered into SPSS 22.0 for descriptive statistics data analysis; 50 test papers were valid. The lowest score in the experimental class was 14, the highest score was 36, and the overall mean score in the experimental class was 23.98. While the lowest score in the control class was 12, the highest score was 33, the overall mean score was 22.98, and the difference between the overall mean scores of the two classes was small. The data analyzed by independent samples t-test, the result of chi-square test, the value of F statistic is 0.004; therefore, the variance of the two classes' performance is chi-square. *T*-test results should be selected, which equal variances assumed (assuming equal variances), with the first line of data as the test result. T-statistic is 0.286, degrees of freedom is 98, and thus above the 0.05 significance level, the null hypothesis is accepted as valid, that is, there is no significant difference between the test results of the experimental and control classes. From the mean of the two classes' scores and the independent samples *t*-test, there is no significant difference between the scores of the experimental class and the control class before the experiment, which can be compared and analyzed. The statistical results of the experimental test are shown in Figure 7.

According to Figure 7, it can be learned that there is very little difference in the number of students using the language blocks in the two classes before the experiment. Comparing the data of the post-test, the number of language blocks in the experimental class was significantly higher than that in the control class: the experimental class used a total of 431 language blocks, with 10.3 blocks per capita, while the control class used a total of 397 language blocks, with 9.7 blocks per capita. In other words, the number of blocks used in the experimental class significantly increased in the post-test, which had a positive effect on the writing performance. According to linguist Ding Yanren's study, it was confirmed that students' writing performance was related to the number of blocks used. Therefore, it can be inferred that the more the number of blocks used by the learners in writing, the higher the writing scores. Thus, it is known that the number of chunks used has a direct effect on writing performance, which means that increasing students' meaningful input of chunks has a positive effect on improving writing performance. Cluster analysis can mine the internal structural features of the data, classify and mark the generated data, and then discover the potential and unknown information in the data.

The model generalization capability was verified, and model strengths and weaknesses were compared on the 17 node network simulation test dataset provided by the NSS subsystem, in which the model version number model9389 had good model generalization capability on the test dataset provided by the NSS subsystem. Meanwhile, by comparing the performance of the model versions model9389 and model9450 on the test set, we can conclude that the log-log loss function is more applicable than the MSE mean square error loss function in the framework of the algorithm model for network transmission delay performance evaluation in this thesis. In the requirement analysis for QISMP inference system management, the tests on personalization management, account management, model version management, model prediction, and history information query, respectively, show that QISMP encapsulates various types of service interfaces to provide network managers with the ability to perform network latency performance evaluation and model version management for special network scenarios. The performance test results of the multidimensional corpus teaching model are shown in Figure 8.

Based on the analysis of the requirements for the simulation packet generation, routing and forwarding, and node

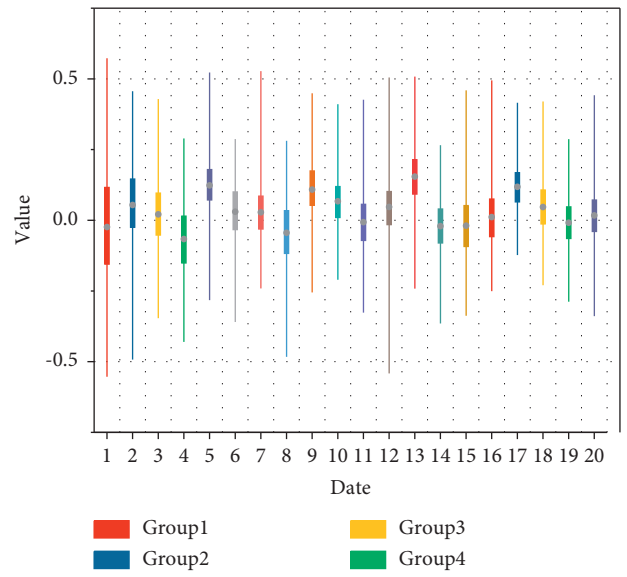


Figure 7: Four groups of control classes' tests with more statistical results.

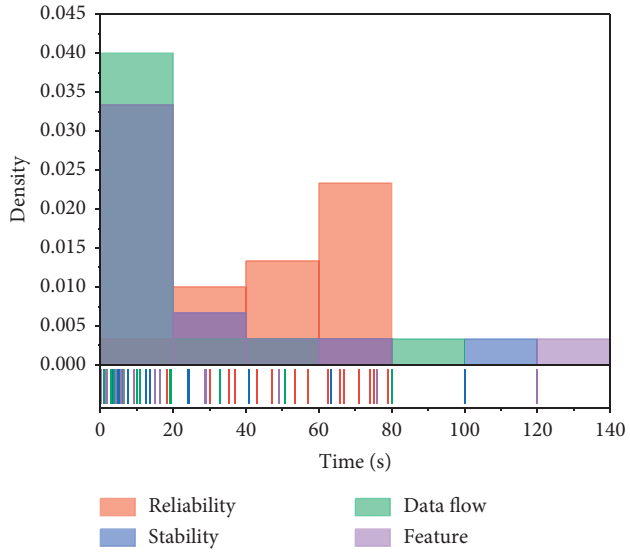


Figure 8: Performance results of the multidimensional corpus English teaching model.

packet management functions that the NSS subsystem network simulation service should have, the test results of the simulation network basic functions under a 14 node network topology with simplified configuration parameters show that the NSS subsystem can complete the transmission performance simulation of specific network scenarios based on the discrete event-driven network simulator OMNeT++. This study provides network simulation data support for the network delay performance modeling by deep learning methods in the NDIS subsystem. The data-driven network QoS inference system based on this thesis is divided into the NSS subsystem, NDIS subsystem, and QISMP. Through the functional tests of the three subsystems, respectively, in this chapter, it can be concluded that the NSS subsystem based on the event-driven network simulator OMNeT++ can effectively provide data support for the network latency performance evaluation model modeling in the NDIS subsystem and the NDIS subsystem. The network delay performance evaluation model NMBGNN designed and implemented in the NDIS subsystem can provide reliable delay prediction service for QISMP, and QISMP can provide model prediction and model version management for network managers with a reasonably designed platform functions.

## 5. Conclusions

In this study, we constructed a data-driven multidimensional corpus-based English teaching model and formed a data-driven teaching and research model. By implementing data-driven teaching and research in university and continuously revising the process and model in practice, we finally built a relatively perfect process model of data-driven teaching and research activities. A framework of transient stable fast batch assessment algorithm is proposed, and the cascaded convolutional neural network is constructed to adaptively select the simulation time window for the samples to be assessed and terminate the time-domain simulation as early as possible while ensuring the accuracy of the assessment conclusion, which is achieved to reduce the overall computational burden of the batch assessment task. In response to the deficiencies in the examination of the relationship between output and input variables, this study explains the basic principles, modeling steps, and advantages and disadvantages of the partial least squares method, introduces the autoregressive latent structure projection algorithm to address the deficiencies of the partial least squares method in the modeling process, and analyzes the modeling principles, steps, and characteristics of the algorithm. In English teaching, teachers are beginning to pay attention to students' sense of experience and to gradually return the classroom center to students. Teachers are also changing their teaching methods and teaching tools to improve teaching standards, following the concept of integrating modern smart teaching technology with English curriculum teaching, further deepening the discussion of combining smart classroom tools with independent learning English skills, and providing new ideas and ways to effectively improve the development of English teaching quality. It provides new ideas and ways to effectively improve the quality of English teaching.

## Data Availability

The data used to support the findings of this study and acknowledgment reference [1] are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by the 2021 Guangxi Higher Education Undergraduate Teaching Reform Project "Research and Practice on Construction of Ideological and Political Evaluation Index System of College English Curriculum Based on CIPP" (project no.: 2021JGB271).

## References

[1] R. Yan, G. Geng, Q. Jiang, and Y. Li, "Fast transient stability batch Assessment using cascaded convolutional neural networks," *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 2802–2813, 2019.

[2] S. C. Silva, T. C. Ferreira, and R. M. S. Ramos, "Data-driven and psycholinguistics-motivated approaches to hate speech detection," *Computación Y Sistemas*, vol. 24, no. 3, pp. 1179–1188, 2020.

[3] M. Mussetta and A. Vartalatis, "Writing across the curriculum in ELT training courses: a proposal using data-driven learning in disciplinary assignments," *International Journal of Teaching and Learning in Higher Education*, vol. 30, no. 2, pp. 300–307, 2018.

[4] I. Ivaska and S. Bernardini, "Constrained language use in Finnish: a corpus-driven approach," *Nordic Journal of Linguistics*, vol. 43, no. 1, pp. 33–57, 2020.

[5] D. Hooshyar, M. Yousefi, and H. Lim, "A systematic review of data-driven approaches in player modeling of educational games," *Artificial Intelligence Review*, vol. 52, no. 3, pp. 1997–2017, 2019.

[6] B. C. Runck, S. Manson, E. Shook, M. Gini, and N. Jordan, "Using word embeddings to generate data-driven human agent decision-making from natural language," *GeoInformatica*, vol. 23, no. 2, pp. 221–242, 2019.

[7] N. I. Khursanov, "On the theoretical and practical foundations of language corpora," *Asian Journal of Multidimensional Research*, vol. 10, no. 9, pp. 311–318, 2021.

[8] J. Jamal, A. Shafqat, and E. Afzal, "Teachers' perceptions of incorporation of corpus-based approach in English language teaching classrooms in Karachi, Pakistan," *Liberal Arts and Social Sciences International Journal (LASSIJ)*, vol. 5, no. 1, pp. 611–629, 2021.

[9] C. Zhang and J. Han, "Multidimensional mining of massive text data," *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 11, no. 2, pp. 1–198, 2019.

[10] M. A. Di Gangi, G. Lo Bosco, and G. Pilato, "Effectiveness of data-driven induction of semantic spaces and traditional classifiers for sarcasm detection," *Natural Language Engineering*, vol. 25, no. 2, pp. 257–285, 2019.

[11] K.-J. Tsai, "Corpora and dictionaries as learning aids: inductive versus deductive approaches to constructing vocabulary knowledge," *Computer Assisted Language Learning*, vol. 32, no. 8, pp. 805–826, 2019.

[12] A. Akbari, "Translation quality research," *Babel. Revue internationale de la traduction/International Journal of Translation*, vol. 64, no. 4, pp. 548–578, 2018.

[13] A. Piotrkowicz, K. Wang, J. Hallam, and V. Dimitrova, "Data-driven exploration of engagement with workplace-based assessment in the clinical skills domain," *International Journal of Artificial Intelligence in Education*, vol. 31, no. 4, pp. 1022–1052, 2021.

[14] D. Hooshyar, M. Yousefi, and H. Lim, "Data-driven approaches to game player modeling," *ACM Computing Surveys*, vol. 50, no. 6, pp. 1–19, 2018.

[15] S. Crossley and M. M. Louwerse, "Multi-dimensional register classification using bigrams," *International Journal of Corpus Linguistics*, vol. 12, no. 4, pp. 453–478, 2007.

[16] A. Batliner, S. Steidl, and C. Hacker, "Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech," *User Modeling and User-Adapted Interaction*, vol. 18, no. 1, pp. 175–206, 2008.

[17] L. Flowerdew, "Applying corpus linguistics to pedagogy," *International Journal of Corpus Linguistics*, vol. 14, no. 3, pp. 393–417, 2009.

[18] A. A. M. Al-Gamal and E. A. M. Ali, "Corpus-based method in language learning and teaching," *International Journal of Research and Analytical Reviews*, vol. 6, no. 2, pp. 473–476, 2019.