International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1–19


https://doi.org/10.70917/ijcisim-2026-0289
Article

# Exploring the Construction of English Translation Corpus and Application of Translation Teaching in Colleges and Universities Based on Parallel Computing

**Weihui Hong***

* College of Education, Jingzhou University, Jinzhou, Hubei, 434020, China; vivien6969@163.com

**Abstract:** The article first collects Chinese-English bilingual corpus through web crawler technology, and then constructs a preliminary Chinese-English bilingual corpus by unifying coding and filtering to improve the quality of the corpus. After that, a distributed system based simple Bayesian algorithm is used to realize large-scale text mining and classification. Then a parallel corpus of English-Chinese translation in colleges and universities is constructed, and a model of the corpus assisting English teaching in colleges and universities is designed, and the practical application effect of the model is analyzed. The results show that: the classification accuracy of this paper's algorithm on the experimental corpus can be stabilized above 95%, and this paper's algorithm works well in the parallel computing performance test, which can straightly keep a lower scale ratio, so that the clusters can be more fully utilized, and has better parallelism. Under the corpus-assisted English teaching practice in colleges and universities, the English proficiency of the students in the experimental class is significantly higher than that of the control class ($p<0.05$), especially in the "vocabulary and grammatical structure, translation, writing" questions, there is a significant difference in the effect between the control class and the experimental class.

**Keywords:** Parallel computing; Distributed system; Simple Bayesian algorithm; Web crawler; English translation corpus

## 1. Introduction

In the context of the new curriculum reform, some college teachers gradually pay attention to cultivating and improving students' translation ability [1]. Students with good translation ability is conducive to their social activities, and they can better understand Western culture, thus forging and improving their knowledge application ability [2-3]. At present, the East and the West have gradually integrated, as the mainstay of the future, Chinese college students need to master a certain degree of English knowledge and translation skills, so as to make due contributions to the country's economic development, which is also an effective embodiment of the development trend in line with the times [4-7].

For this reason, the literature [8] emphasizes the importance of English translation teaching, which can promote students to identify the differences in structure and vocabulary, enhance their comprehension ability and form their own way of thinking. Literature [9] analyzes the role of improving students' translation skills in promoting "social change competence", and indicates that translation skills are an important way for students to promote social justice and social empowerment. Literature [10] discusses the importance of the impact of translation competence on day-to-day business operations and aims to analyze the benefits of utilizing translation activities in the development of language skills for learners of English for Specialized Purposes (ELLs) in order to meet the needs of the labor market. Literature [11] examines a team translation program on Aboriginal culture, aiming to enhance the intercultural communication skills and understanding of university

students by having interactive team translation activities. Literature [12] emphasizes the important role played by translation in the mastery and application of English skills, based on which it examines how to improve the quality of translation teaching for English majors in colleges and universities and proposes measures to optimize their translation teaching methods.

Literature [13] points out the negative situation of teachers and students in English translation teaching through the results of a survey of English translation teachers and students, and proposes a model for applying the principles of positive psychology to the English classroom, aiming to cultivate students' positive attitudes, adaptability and thinking in translation teaching. Literature [14] points out that there are problems such as the lack of independent translation courses, high-quality teaching materials and qualified teachers for teaching translation to non-English majors in China at present, and proposes corresponding countermeasures. Literature [15] describes the role of translation teaching in cultivating translators, and points out that it has long existed problems such as weak teachers and outdated teaching modes, and that colleges and universities must adopt strategies to improve the quality of translation teaching. Literature [16] explains that the current English translation teaching mainly adopts the teaching method based on teaching materials and single form, in order to improve the teaching effect, the strategy to improve the effect of English translation teaching is discussed, and the research shows that the teaching method of stimulating students' interest and using modern equipment helps to cultivate students' interest in English translation, thus improving the effect of translation teaching. Literature [17] points out the shortcomings of English majors in Chinese colleges and universities in actual translation work, and analyzes the linguistic differences between English and Chinese, the lack of necessary translation skills and other problems, and puts forward strategies such as strengthening cultural learning and teaching translation skills. To sum up, the current translation teaching in colleges and universities is generally faced with such problems as the amount of teaching hours is small, the content of teaching materials is single and lagging behind, the classroom teaching mode is obsolete, and the evaluation system is not reasonable enough, etc. The translation teaching mode needs to actively implement reforms, change the concept of teaching and the understanding of teaching, and innovate the mode of teaching, and the construction of an English translation corpus promotes the realization of this goal [18-22].

Translation corpus is very advantageous for the study of translation students in colleges and universities, which not only contains rich online translation resources and textbook translations, but also retains the records of every translation practice of students, which can be used repeatedly to find students' knowledge weaknesses and carry out targeted graded and modular intensive practice, focusing on breaking through the bottlenecks in the learning process [23-27]. The translation corpus contains finance, history, art, law and other aspects of the content, the source of the text material, such as celebrity speeches, large-scale conferences, newspapers and magazines, etc., to meet the different translation level of students to practice [28-29]. When practicing with the translation corpus, students have more freedom, they can do it at any time after class, and understand their own learning level by virtue of the exercises given by the system, forming a record with durability, and it is also more convenient for teachers to understand the students' practice [30-32]. In the corpus of exercises can automatically help students to correct errors and verification, students' after-school practice is no longer blind, quickly realize the translation of the content of the Q&A, and according to the error-prone points to expand their knowledge, students can better master the relevant knowledge [33-36].

Regarding the research on the application of translation corpus in translation teaching, literature [37] explored the impact of translation corpus on students' English translation learning, and the questionnaire showed that most students welcomed the corpus because it effectively improved their translation skills. Literature [38] describes that corpora have been widely used in translation teaching and learning, and based on the literature review emphasizes the positive impacts that corpora bring in the translation classroom. Literature [39] examines the effectiveness of corpus-based pedagogies and other advantages they bring to translation teaching, aiming to inspire teachers to use corpus-based pedagogies in translation teaching. Literature [40] describes the strategies used by students when translating and introduces parallel corpora as a reference, emphasizing that the use of corpora in the translation classroom allows students to access a great deal of information about language use, which helps to improve their translation efficiency. Literature [41] emphasizes that corpora are reference databases of language and that they can play an important role in the exploration of translation solutions for bilingual parallel texts, and in their regular use in professional texts. Literature [42] presents books on teaching translation with corpora, which comprehensively present those key approaches that rely on corpus applications in translation teaching. Literature [43] introduces parallel translation corpora, pointing out that they can lead to a learning experience with personalization, but at the same time, they suffer from poor data quality, privacy concerns, and inability to scale. However, the existing research mainly focuses on a single application scenario of the corpus, which cannot realize

the efficient demand in complexity corpus processing.

In this paper, we first collect the corpus of English translations in some bilingual websites through python web crawler technology, and unify the coding and filtering of the corpus to get a preliminary Chinese-English bilingual corpus. After that, a simple Bayesian classification algorithm is introduced to solve the problem of text categorization of the corpus. However, since the algorithm cannot be directly used to mine large-scale text on cloud computing, this paper adopts Spark distributed system to realize the parallelization of the plain Bayesian algorithm, and analyzes the performance of this optimized algorithm. Finally, from the perspective of the application of English translation teaching in colleges and universities, a corpus-assisted college English teaching model is designed, and a case study is conducted to analyze the effect of the application of this model in English teaching in colleges and universities.

## 2. English Translation Corpus Classification Method Based on Parallel Computing

### *2.1. Chinese-English parallel corpus collection and preprocessing*

#### 2.1.1. Principles of web crawlers

A web crawler [44] is actually a simple program or script that focuses on traversing the content of an entire website through a given web address and then following a specified strategy. When it encounters data that meets the requirements, it downloads it and stores it in a local folder, and when it encounters unwanted data, it skips it and then continues the traversal to find the next data.

#### 2.1.2. Algorithms for crawling Chinese and English data

This section focuses on the collection of bilingual corpus, so it is necessary to find the first address of the website in both Chinese and English. Next, we need to analyze the HTML structure of the first address of the website and collect the chapter addresses inside, and then collect the desired corpus through the web crawler system. The specific steps of the web crawler algorithm are as follows:

Step 1: First of all, we have to analyze the characteristics of Chinese and English bilingual websites to design the crawler system, especially we have to have a certain understanding of the structure of Chinese and English bilingual websites, choose according to the Chinese and English materials that need to be collected, and then organize the websites that need to collect and corpus. The web addresses collected this time are mainly those of the Chinese-English dual websites, and the first address of the Chinese-English website is initially selected, and the Chinese-English chapter addresses are collected through the first address of the website and stored in a txt file after collection.

Step 2: Next, we need to carefully read the HTML structure of the website in order to collect the needed content. Since we are going to collect the Chinese-English bilingual corpus, we need to understand the characteristics of the Chinese-English material distribution of this kind of website. Then according to these characteristics to design the crawler, mainly the Chinese and English will be divided into two crawlers to obtain.

Step 3: The next step is to collect according to the paragraphs in the chapter, because if we collect according to the whole article, it is easy to collect some useless information, so we need to collect the body of the article.

Step 4: After collecting a corpus, we need to name the collected corpus, mainly according to the title of each article, and complete the collection of all the websites in the txt in this order, which basically completes the collection of the corpus.

#### 2.1.3. Corpus preprocessing

Before preprocessing the corpus, the first thing that needs to be done is to integrate the collected corpus, which was originally collected according to the chapters, now all the corpora need to be synthesized into a single text, which only needs to be merged according to the order in which it was collected by the crawler. After integration, the corpus is preprocessed, which is mainly divided into two steps, the first step is to standardize the encoding of the collected text. The second step is to filter some titles and network tags etc. in the text, mainly using regular expressions to realize text filtering, these two steps are mainly to improve the quality of the collected corpus.

(1) Unified Coding

Text character encoding is normalized. At present, the encoding format of the corpus is mainly the

utf-8 format, and the use of other formats will affect the reading of the corpus in the application.

(2) Filtering text

Next, we will use regular expressions to filter and delete the corpus collected in the previous section, which is mainly used to delete some unwanted contents. Regular expression can be a combination of various types of characters, there is no fixed combination of ways, mainly according to what you want to filter, to construct the regular expression formula.

## *2.2. Spark-based parallelization of plain Bayesian corpus classification*

### 2.2.1. Plain Bayesian classification

(1) Bayesian

The Bayesian formula is a statistical principle that combines a priori knowledge of classes with new evidence gathered from data. The four Bayesian classification algorithms that are currently being studied are Naïve Bayes, TAN, BAN, and GBN.

Among several of the most important formulas in probability theory are Bayesian formulas - used to describe the relationship between two conditional probabilities, such as $P(A \mid B)$ and $P(B \mid A)$. The formula is represented as follows:

$$P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A)} \tag{1}$$

where $P(B)$ is called the prior probability of $B$ and $P(B \mid A)$ is called the posterior probability of $A$ conditional on $B$. Similarly, $P(A)$ is called the prior probability of $A$ and $P(A \mid B)$ is called the posterior probability of $B$ conditional on $A$.

(2) Plain Bayes

Plain Bayesian classification [45] is a probabilistic statistical classification algorithm based on the Bayesian formulation and the assumption of conditional independence of features. Its principle is very simple: for a certain prediction item, the probability of the prediction item for each classification is calculated separately, and then the classification with the largest probability is selected as its prediction classification.

(3) Parameter estimation and event modeling

I. Polynomial model

Polynomial Simple Bayes, which is one of the two classical Simple Bayes variants used for text classification, views a document as a series of ordered feature words, and the set of ordered feature words is represented by the vector $x = \{a_1, a_2, \ldots, a_m\}$, where the values of the elements in the vector are represented by the TFIDF values, and $m$ is the number of feature words. The document is thus represented as a vector $x$.

In the polynomial model, word frequencies are computed on the training set with maximum likelihood estimation to estimate the class conditional probability $P(a_j \mid y_i)$:

$$\hat{P}(a_j \mid y_i) = \frac{N_{y_i a_j}}{N_{y_i}} \tag{2}$$

where $N_{y_i a_j} = \sum tf(a_j \mid y_i)$ denotes the word frequency of the training set samples belonging to class $y_i$ that contain feature word $a_j$, and $N_{y_i}$ denotes the sum of the word frequencies of all feature words belonging to the training set samples of class $y_i$.

If a given class and feature word do not co-exist in the training data, then the probability estimation based on word frequency is 0 i.e. $P(a_i \mid y_j) = 0$, for such a case equal to 0, it is better to add the value of its count to $\lambda(\lambda \geq 0)$, and the above equation is converted to:

$$\hat{P}(a_j \mid y_i) = \frac{N_{y_i a_j} + \lambda}{N_{y_i} + \lambda n} \tag{3}$$

where $n$ is the number of classes. When $\lambda = 1$ is set, it is called Laplace smoothing, and when $\lambda < 1$ is set, it is called Lidstone smoothing. If multiplied by many conditional probability values, this can lead to floating point overflow. Since $\log(xy) = \log(x) + \log(y)$ , it is then possible to turn a consecutive multiplication into a consecutive addition by means of the log function of probability:

$$\log\left(P\left(x \mid y_i\right)\right) = \log\left(\hat{P}\left(y_i\right)\prod_{j=1}^{m}\hat{P}\left(a_j \mid y_i\right)\right) = \log\hat{P}\left(y_i\right) + \sum_{j=1}^{m}\log\hat{P}\left(a_j \mid y_i\right) \tag{4}$$

The plain Bayesian classifier formulation is improved by using weighted TFIDF weights calculation:

$$P\left(x \mid y_i\right)P\left(y_i\right) = P\left(y_i\right)\prod_{j=1}^{m}P\left(a_j \mid y_i\right)^{n_j} \tag{5}$$

where $n_j$ is the word frequency of the $j$ th feature.

According to the above equation, the text polynomial plain Bayesian classifier becomes a linear classifier:

$$\log\left(P\left(x \mid y_i\right)P\left(y_i\right)\right) = \log\hat{P}\left(y_i\right) + \sum_{j=1}^{m}n_j \cdot \log\hat{P}\left(a_j \mid y_i\right) = b + w_j^T \cdot x \tag{6}$$

where $b = \log\left(\hat{P}\left(y_i\right)\right)$ and $w_j = \log\left(\hat{P}\left(a_j \mid y_i\right)\right), x = \{n_1, n_2, \ldots, n_m\}$ for the text to be tested.

II. Bernoulli model

In the multivariate Bernoulli event model, the feature terms are described in terms of Boolean variables that are assumed to be independent. The feature vector $x\left(a_1, a_2, \ldots, a_m\right)$ of the document has dimension $m$ , where is the number of feature words in the whole dictionary, and 1 and 0 are used to denote whether or not the feature word occurs in the document, respectively; this representation also does not take into account the order and the number of times that the feature word occurs in the document, because the feature vector $x$ of the events are generated by heavy Bernoulli experiments, so the formula can be written as:

$$P\left(x \mid y_i\right) = \prod_{j=1}^{m}\left(bP\left(a_j \mid y_i\right) + (1-b)\left(1 - p\left(a_j \mid y_i\right)\right)\right) \tag{7}$$

where $b \in \{0,1\}$ ; $P\left(x \mid y_i\right)$ denotes that the conditional probability of a document belonging to the class $y_i$ is the product of the class-conditional probabilities of all the feature terms, with:

$$\hat{P}\left(a_j \mid y_i\right) = \frac{n_{ji}}{n_i} \tag{8}$$

where $n_{ji}$ is the number of documents containing the feature term $a_j$ in the class $y_i$ in the training set, and $n_i$ is the number of all documents in the class $y_i$ . A smoothing factor is introduced to address the case where the class conditional probability is equal to 0, i.e., an estimation method is used to estimate the conditional probability:

$$\hat{P}\left(a_j \mid y_i\right) = \frac{n_{ji} + mp}{n_i + m} \tag{9}$$

where $m$ refers to the equivalence sample size and $p$ is specified by the user. The commonly taken parameter values are $= 2, p = 1/2$ .

### 2.2.2. Feasibility Analysis of Parallelizing the Plain Bayesian Algorithm

The Analytic Bayesian algorithm consists of two main parts, the two processes of classifier training and prediction. The feasibility of parallelization of these two processes is analyzed separately below.

(1) Parallelization of the classifier training process

The training process mainly includes the computation of word frequency, class number, class prior probability and class conditional probability based on the training samples. The training set can be divided into multiple partitions, in which the training process of each text is the same and independent of each other. The granularity of partitioning can be set according to the number of computing nodes in the cluster, and the smallest partition is each document.

(2) Parallelization of classifier prediction process

The essence of the prediction process is to calculate the class conditional probability of each feature word and sum it up to get the probability that the document belongs to each class and take the maximum value of the class as the classification result. The most time-consuming part of the process lies in the large number of computations required, while the computation of the class conditional probability of each feature word is independent of each other, so the process can be executed in parallel, by the computational nodes to complete the computation of the local partition, and finally summarize the results of the intermediate calculations to obtain the classification results.

### 2.2.3. Parallelization of Hadoop-based Plain Bayesian Text Classification

In this section, the Hadoop distributed system [46] is used to implement the parallelization of the plain Bayesian algorithm, which provides the storage capacity of HDFS and the computational power of MapReduce to support the implementation of parallel class algorithms.

(1) Parallelization of classifier training process

Because it is implemented based on Hadoop, whose programming model is classical MapReduce, the classifier training is realized by two sequential combinatorial MapReduce jobs, in which the output of the former MapReduce is used as the input of the latter one, and the process of parallelizing the classifier training is shown in Fig. 1.

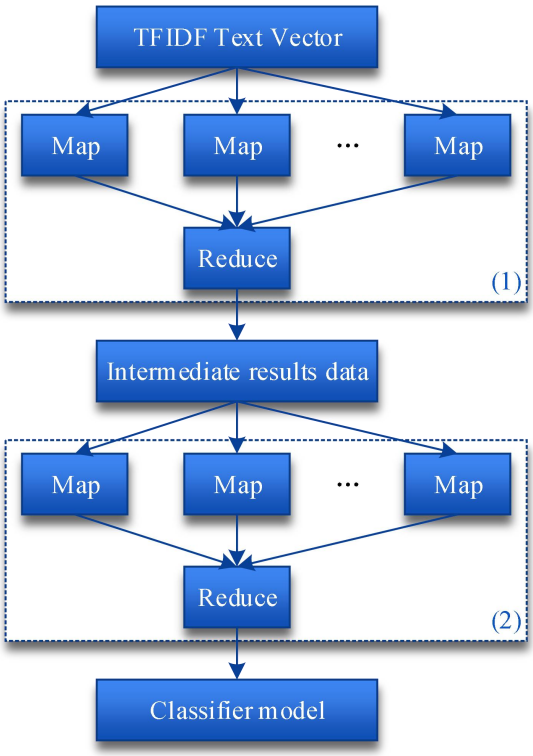


***Figure 1.*** Parallelization of classifier training

(2) Parallelization of the classifier prediction process

The classifier prediction process is relatively simple and is realized with a MapReduce job. First, each Map receives the set of test text vectors, calls the classification model obtained from the training

process, and computes the comparison to derive the class labels of the test documents. Second, Reduce merges the results computed by each Map and derives the test set accuracy.

### 2.2.4. Parallelization of Spark-based Plain Bayesian Text Classification

Similar to the Hadoop-based parallelization in the previous section, the Spark-based implementation is also divided into two processes to be analyzed: training and prediction, with the difference that the Spark-based implementation focuses on the algorithmic design changes around the RDD.

(1) Parallelization of classifier training process

After the text vectorization representation is completed, the current task is to build a classification model using the training samples that have been labeled with categories. We represent the training samples in the form of key-value pairs where each row is a class label and a feature vector stored on HDFS.

(2) Parallelization of the classifier prediction process

Once the training process is completed, the Naive Bayes classification model is obtained. The next step is to use the established model to predict the classification: first, read the test samples to form the RDD, and then through the map function according to the trained model for the text vector inside the RDD, calculate the probability that the test samples belong to each class, and take the class with the largest probability as the class number. Finally, the classification results of all test samples are saved to local disk or HDFS using the savsAsTextFile method provided by Spark.

## *2.3. Analysis of the results of the application of the algorithm in this paper*

### 2.3.1. Classification Performance Testing and Analysis

This paper adopts check accuracy rate, check completeness rate and F1 value as evaluation indexes, and in parallel performance evaluation text adopts acceleration ratio, scale growth, scalability as evaluation indexes, and compares the running time and acceleration ratio of Spark and MapReduce.

The corpus selected for this experiment is the Chinese corpus of Q University, and all the data of the six most representative categories are selected as the input text of a total of 15,000 documents, including 10,000 documents in the training set and 5,000 documents in the test set.

During the experiment, the classification effect of the algorithm in this paper is compared and analyzed with that of KNN, Simple Bayes, Decision Tree C45 and traditional Random Forest (RF) algorithm.

The results of the comparison of the check accuracy rate of different classification algorithms are shown in Table 1. The comparison results of the check accuracy rate of different classification algorithms are shown in Table 2. It can be seen that the KNN algorithm has the worst classification effect on this corpus, and the accuracy and recall are low on most categories; the overall situation of the plain Bayesian algorithm is improved compared with the KNN algorithm, and the accuracy and recall are more balanced, but the overall effect performance is average, and the accuracy and recall of all categories fluctuate between 76.09% and 87.85%; the C45 decision tree has a better ability to classify this corpus C45 decision tree has a better ability to classify this corpus, and the accuracy is stable at about 85%; RF algorithm is better than the decision tree C45 algorithm on the whole, but the overall effect is not as good as this paper's algorithm. Comparison shows that this paper's algorithm has the best classification ability for this corpus, and the classification effect is the best in all categories, and its classification effect among all categories has no big deviation, which is stable at more than 95%, and the accuracy and recall rate are relatively balanced.

***Table 1.*** Comparison of precision rates for different classification algorithms

| **Survey details** | **KNN** | **NB** | **C45** | **RF** | **Spark-based Naive Bayes** |
|---|---|---|---|---|---|
| | **Precision ratio(%)** | | | | |
| Physical culture | 87.13 | 86.86 | 85.73 | 87.97 | 98.14 |
| Politics | 80.88 | 78.12 | 85.02 | 87.41 | 95.81 |
| Environment | 82.09 | 84.06 | 86.99 | 89.03 | 99.27 |
| Economy | 84.92 | 76.09 | 82.06 | 82.79 | 97.06 |
| Agriculture | 81.98 | 77.39 | 85.18 | 84.34 | 96.05 |
| Computer | 63.89 | 87.85 | 88.91 | 88.73 | 98.93 |
| Average value | 80.15 | 81.73 | 85.65 | 86.71 | 97.54 |

| | | | | | |
|---|---|---|---|---|---|
| F1 | 73.91 | 80.15 | 84.95 | 85.03 | 96.91 |

***Table 2.*** Comparison of recall rates for different classification algorithms

| Survey details | KNN | NB | C45 | RF | Spark-based Naive Bayes |
|---|---|---|---|---|---|
| | Recall ratio(%) | | | | |
| Physical culture | 68.77 | 80.15 | 84.05 | 86.79 | 98.43 |
| Politics | 66.73 | 80.05 | 80.05 | 83.59 | 94.67 |
| Environment | 68.02 | 79.13 | 88.12 | 87.6 | 97.79 |
| Economy | 69.16 | 76.29 | 81.84 | 83.67 | 96.99 |
| Agriculture | 69.93 | 76.2 | 84.15 | 86.15 | 97.98 |
| Computer | 82.02 | 81.92 | 87.08 | 97.25 | 99.04 |
| Average value | 70.77 | 78.96 | 84.22 | 87.51 | 97.48 |

Experiments were conducted using Random Forest Algorithm (RF) and Random Forest Algorithm based on Rough Set Theory (R-RF), Random Forest Algorithm with Weighted Voter (V-RF) and this paper's algorithm to analyze the effect of the improvement of rough set theory based on rough set theory in classification performance.

The results of the comparison experiments of this paper's algorithm are shown in Figure 2. The experiments show that compared with the RF algorithm, the V-RF algorithm has deteriorated in local performance, but the overall performance has been slightly improved, mainly because the addition of the weighting factor has solved the situation that the voter casts more than one class. Compared with the RF algorithm, the R-RF algorithm has worse local performance, but the overall performance has a small increase, mainly because the random subspace based on rough set theory, strengthen the weaker decision tree classification ability due to randomness, compared with the rest of the algorithms, this paper's algorithm to improve the effect is more obvious. Due to the overall distribution of randomness is relatively stable, the voter is not a large number of votes to multiple classes, and the random subspace overall tends to be stable, but this paper's algorithm still has a 1.26% to 6.09% improvement in the F1 value.

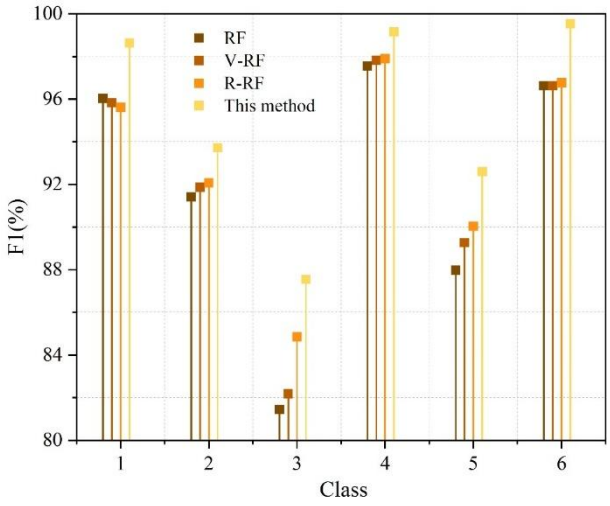


***Figure 2.*** Random forest algorithm comparison results

The experiment also investigates the effect of random forest size on the classification results: the changes of F1 values of each class when the number of decision trees increases are shown in Fig. 3. The results show that the overall classification performance of the corpus tends to stabilize when the number of decision trees reaches 30, but there is still room for some classes to rise, and finally the classification performance of each class tends to converge when the number of decision trees is 70. This is because the decision tree in the random forest contains only a small part of the attributes. When the size of the random forest is small, the classification model can not fully reflect the training set, at

this time part of the performance of the class is poor; when the random forest reaches a certain size, can fully reflect the training set, the classification performance will tend to stabilize.

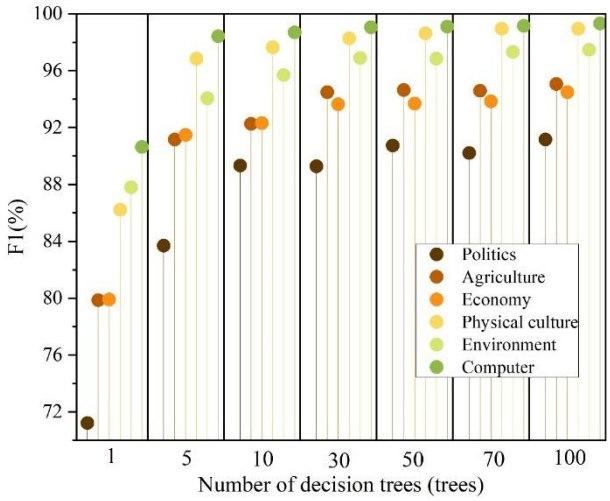


***Figure 3.*** How various F1 values change as the number of decision trees increases

## 2.3.2. Parallel Computing Performance Testing and Analysis

In a cluster of 16 nodes, the original corpus is enlarged a number of times as input data, and the text is merged to be about 3.5GB in size, and the experiments are conducted using Speedup, Scaleup, and Sizeup to evaluate the parallel performance.

(1) Speedup ratio test and analysis

The speedup ratio reacts to the improvement of parallel performance with the increase of the number of nodes while keeping the data unchanged. The acceleration ratio test results are shown in Fig. 4. Under the experimental conditions, the acceleration ratio of this paper's algorithm is lower than the ideal linear acceleration ratio due to the loss of inter-node communication, with an overall decreasing trend, which is caused by the fact that with the expansion of nodes, it makes the communication time between the nodes accounted for a large proportion.

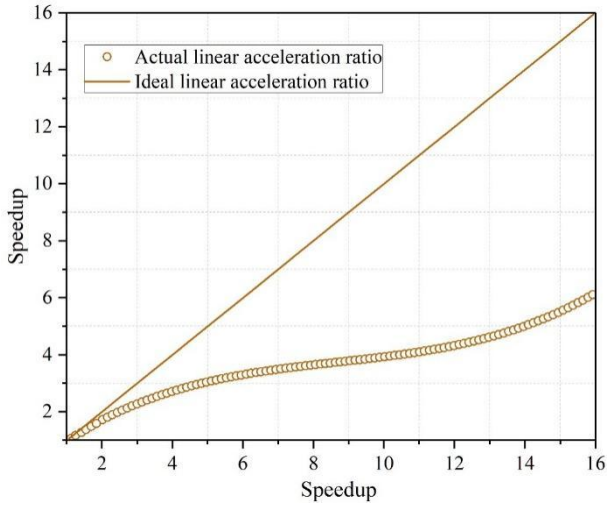


***Figure 4.*** Speed up test results

(2) Scale ratio test and analysis

The scale ratio reacts to the parallel performance of the algorithm itself, keeping the number of nodes unchanged and increasing the data size during the experiment. The results of the scale ratio test are shown in Figure 5. Under the experimental conditions, the algorithm in this paper has kept a low scale ratio. There is a rising trend with the increase of nodes. This is because increasing the data scale while the nodes remain unchanged makes the cluster more fully utilized, indicating that the algorithm in this paper has better parallelism.

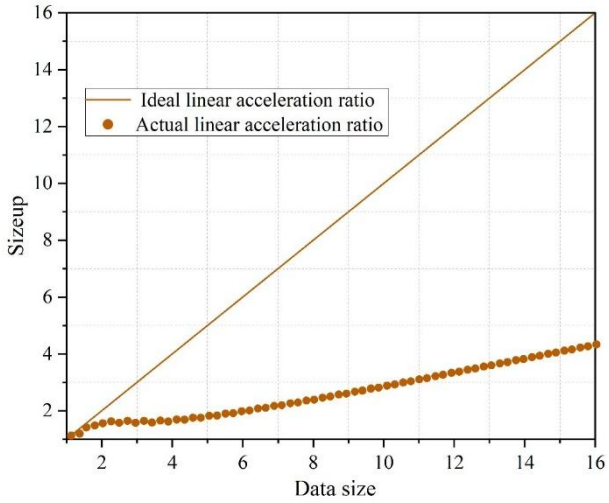


***Figure 5.*** Scale comparison results

(3) Scalability test and analysis

Scalability reacts to the performance of the algorithm for data expansion, increasing the number of nodes while increasing the data in the experiment. The scalability test results are shown in Figure 6. Under the experimental conditions, the scalability of this paper's algorithm shows a decreasing trend and is lower when there are more nodes, which is due to the increase in running time caused by the increase in the overhead of inter-node communication with the increase of cluster nodes within a certain range.

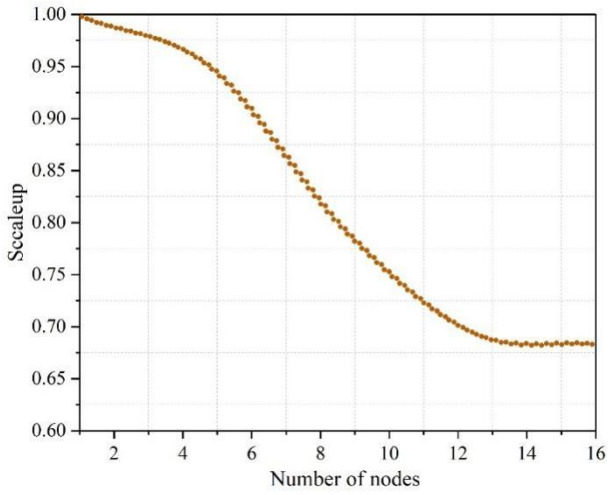


***Figure 6.*** Scalability test results

(4) Comparison of running performance under Spark and MapReduce frameworks

In order to illustrate the efficiency of the Spark-based in-memory computing framework, the experiment kept the data unchanged and counted the SparkReduce running time. The experiments analyze the running time of MapReduce and Spark under multiple nodes.The results of the comparison of the running performance of Spark and MapReduce framework are shown in Figure 7. The experiments show that the performance improvement of Spark over MapReduce is larger when there are fewer nodes. With 1 node, MapReduce runs in 28.27 times the time of Spark. As the nodes increase and the cluster is fully utilized this gap decreases and with 16 nodes, the running time of MapReduce is 12.15 times that of Spark. Therefore, the algorithm of this paper in Spark framework has higher efficiency compared to MapReduce.

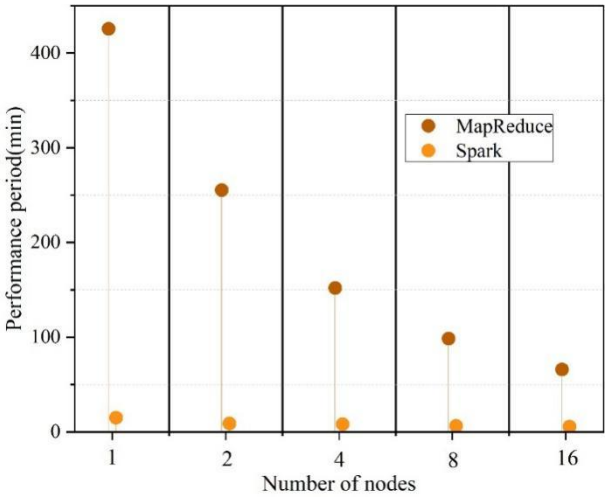


***Figure 7.*** Spark & MapReduce framework performance comparison results

In order to more fully investigate the parallelism of this paper's algorithm under the two frameworks, the experiment also compares the acceleration ratio of this paper's algorithm on MapReduce and on Spark. The results of the parallelism comparison of the speedup ratio of this paper's algorithm under the two frameworks are shown in Figure 8. The experiment shows that the trend of the acceleration ratio of the two is the same, but the acceleration ratio of MapReduce is closer to linear, while the acceleration ratio of Spark is relatively low, which is caused by the fact that on Spark, the running time of this paper's algorithm is shorter, which makes the node communication time account for a larger proportion.

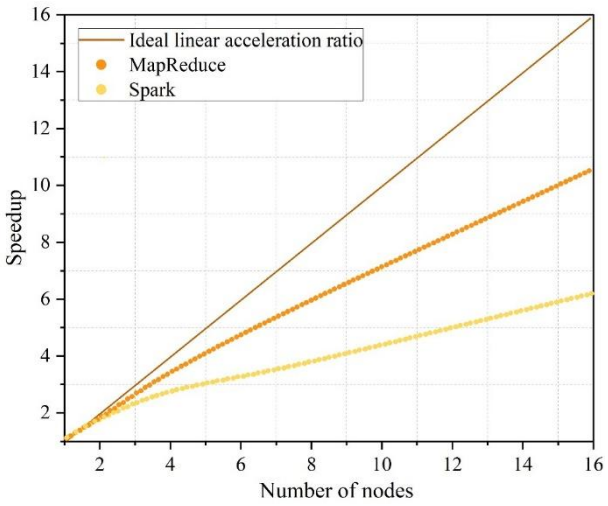


***Figure 8.*** Comparison of acceleration versus parallelism under two frameworks

## 3. Case Study of English Translation Corpus Construction and Teaching Application in Colleges and Universities

### *3.1. Construction of a Parallel Corpus of English-Chinese Translation in Colleges and Universities*

The functional structure of the English-Chinese translation parallel corpus system is shown in Figure 9. The system contains seven functional modules: crawler module, sentence alignment module, corpus import module, corpus browsing and modification module, sentence pair retrieval module, corpus export module, and human-computer interaction module.

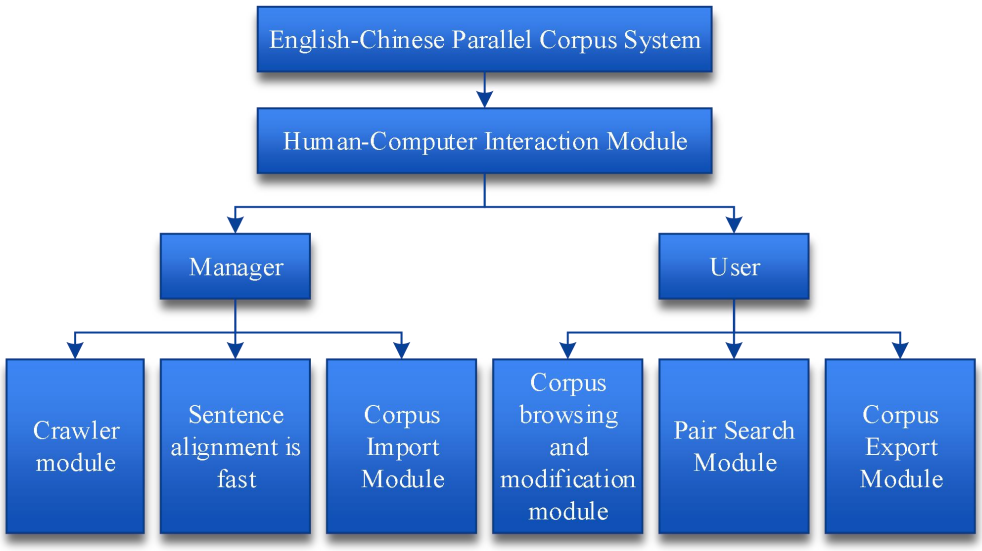


***Figure 9.*** The functional structure of the parallel corpus system

(1) Crawler module

The crawler module realizes the function of downloading English-Chinese bilingual archives, which uses Python's selenium module to monitor Google Chrome, and connects to the Pyspider crawler framework running on port 5000, Pyspider will create a new project according to the URL and customized project name input by the user, and then After that, selenium will automatically modify the Pyspider crawler code, save the project and run the project, and finally output the English-Chinese bilingual file in JSON format.

(2) Sentence Alignment Module

Sentence Alignment Module realizes sentence-level alignment of English-Chinese bilingual files. Users need to manually convert the English-Chinese bilingual files obtained by the crawler module from JSON format to TXT format, and alternate the English and Chinese chapters that are translated into each other. Input the processed TXT document to the sentence alignment module, the module will automatically complete the chapter and paragraph level alignment of the English-Chinese bilingual corpus, and then use the algorithm to realize the sentence level alignment on the basis of the paragraph level alignment of the English-Chinese parallel corpus. Finally, the sentence alignment module will output the aligned sentence pairs and display the alignment results on the graphical interface for users to review.

(3) Corpus Import Module

The corpus import module will store the aligned sentence pairs output by the sentence alignment module into the database to form a corpus. MYSQL is chosen as the database, and there are two roles for users and administrators. The user role can read and call the corpus in the database, but cannot add, modify or delete the database. The administrator role can complete the modification of the database.

(4) Database browsing and modification module

The database browsing and modification module provides an interface for users to browse all the English-Chinese parallel pairs stored in the corpus through a graphical interface, and allows users to modify the contents of the pairs or delete some pairs directly. This module realizes the user's modification and deletion functions of the English-Chinese parallel corpus, and the user can manually modify some observed translation errors or directly delete some sentence pairs with sentence alignment errors in the process of using the English-Chinese parallel corpus, which helps to improve the sentence alignment accuracy of the English-Chinese parallel corpus and make the corpus more reliable.

(5) Sentence Pair Retrieval Module

The Sentence Pair Retrieval Module provides an interface that allows users to retrieve English-Chinese parallel pairs from the English-Chinese Parallel Corpus. Users can retrieve pairs by the serial number of the pairs, and all eligible pairs will be displayed on the user's graphical interface after inputting the serial number of the pairs. Users can also use English words and Chinese words to retrieve sentence pairs. Entering English or Chinese keywords, the user GUI will display all sentence pairs containing the keywords.

(6) Corpus Export Module

The corpus export module provides an interface to allow users to export the corpus from the

English-Chinese parallel corpus, and the format of the exported file can be selected between TXT format and CSV format. Users can export the English-Chinese parallel corpus to meet their own needs through this module.

(7) Human-Computer Interaction Module

The human-computer interaction module manages the user graphical interface and presents all interfaces for users. Users can call all the functions of other modules here. The user graphical interface is completed written using Python's Tkinter module.

## *3.2. Design of Corpus-Assisted English Teaching Model in Colleges and Universities*

The project tries to introduce the corpus method into four specific teaching aspects of vocabulary, grammar, translation, writing and vocabulary of English in colleges and universities. On the one hand, it improves students' language awareness and contextual understanding through the existing large-scale corpus, and on the other hand, it collects students' typical errors through the self-built small-scale corpus in order to improve the accuracy of the foreign language application.

### 3.2.1. Corpus-assisted vocabulary instruction

The focus of vocabulary teaching classroom lectures is the conceptual meaning of words, and the vocabulary exercises designed are also English-Chinese sentence translation or sentence construction centered on the conceptual meaning, neglecting the teaching of word collocations, the contexts in which the words appear and their connotative meanings. Teachers can teach vocabulary based on the corpus, using parallel corpus in the opus corpus, searching for words to be learned using paraconc, expanding different meanings and lexemes of the words, and verifying the common usage of the words.

### 3.2.2. Corpus-assisted grammar instruction

In the teaching of English grammar in colleges and universities, teachers can retrieve applicable example sentences from the American Corpus of Contemporary English (coca) and the British National Corpus (bnc) to show students the corresponding grammatical laws with real corpus, so as to let the students improve their linguistic communication skills while mastering grammatical knowledge.

### 3.2.3. Corpus-assisted teaching of translation

Because of its own characteristics, the bilingual corpus is very suitable for translation teaching, which on the one hand can provide translation examples and improve translation efficiency; on the other hand, it can objectively assess the quality of students' translations and help students improve their translation level. In the practice of translation teaching, tmxmall corpus platform can be used, the advantage of the platform is that in addition to providing bilingual parallel search, it can also provide text alignment and other functions, which can convert the translated content into parallel corpus in time and save it, in the form of translation memory.

### 3.2.4. Corpus-assisted writing instruction

Incorporating a corpus into English writing is effective in improving students' writing skills. A corpus allows students to have direct contact with a large and comprehensive corpus, which is more intuitive and easier to understand than a concise and generalized textbook. Students are equivalent to being in a real language environment, and can actively discover the rules of language use instead of passively inputting language rules. Currently, the writing platform of “Critique.com” has developed an automatic essay review system based on corpus and cloud technology, which reduces teachers' burden of reviewing to a certain extent, but the reliability of reviewing the content and structure of essays is not high, therefore, teachers can establish a small-scale corpus based on the teaching class by combining with the functions of “Critique.com” in their teaching practice, and analyze in depth the content of essays. Therefore, teachers can build a small corpus based on the teaching class by combining some functions of the Critique and Reform Network, and analyze the writing problems of the students in the class in depth. Based on the above reforms, the design of the corpus-assisted English teaching process in colleges and universities is shown in Figure 10.

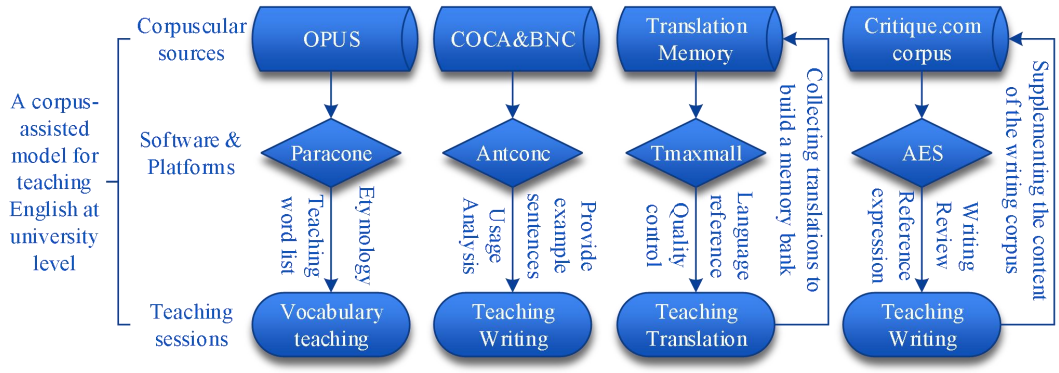


***Figure 10.*** The teaching process of college English with corpus assistance

## *3.3. Research Design of Corpus-Assisted English Teaching and Learning in Colleges and Universities*

### 3.3.1. Research questions

This study mainly answers the following two questions:

(1) Can corpus-assisted instruction improve the English proficiency of college students?

(2) How can corpus be effectively utilized to assist the teaching of English reading and writing in colleges and universities?

### 3.3.2. Objects of study

This study was conducted on two naturalistic teaching classes in the first year of college M. There were 50 students in each of the two classes. The university has added reading and subsequent writing questions to the English performance assessment In addition, the students in these two classes have comparable English proficiency, the same instructor, and the same amount of weekly class time, which maximizes the experimental reliability.

### 3.3.3. Research methodology

The main methods used in this study are experimental method, questionnaire method, test method and interview method. Through the experimental method and test method, this paper collects the data related to the performance of reading and subsequent writing before and after the experiment. Through the questionnaire method and interview method, students' attitudes and suggestions about corpus-assisted reading and writing instruction are collected.

The questionnaire mainly investigated the students from the following perspectives: whether the students felt that the corpus-assisted reading and writing mode had improved their self-reading and writing ability; how the students recognized the corpus-assisted reading and writing mode; whether the students would continue to use the corpus-assisted reading and writing practice in their future studies. To ensure the validity of the questionnaire results, the questionnaire was conducted anonymously.

### 3.3.4. Research process

In the experimental process, the control class used the traditional reading followed by writing teaching model, and the teaching process was divided into the following three steps:

Step1: Reading and then writing the original text In this process, students first read and then write the original text independently. Teachers do not explain the reading and writing texts, and students read the texts independently and silently to find out the time, place, characters, and basic information such as the cause, the passage, the conflict, the climax, and other basic information about the development of the storyline.

Step2: Writing Students make reasonable speculations based on the first sentences of the given two paragraphs and finish the sequel in 25-30 minutes.

Step3: The teacher will lead the students to find out the core elements of the story (when, where, who, what, why and how), analyze the storyline, and find out the cause, passage and conflict of the story. Then the teacher scores the students' work according to the scoring criteria of reading and writing. Finally, the excellent model essays were analyzed to highlight what was worth learning. In the course of the experiment, the instructional design of the experimental class and the control class was quite

different. The experimental class adopts the corpus-assisted reading and subsequent writing teaching model, and this teaching process is divided into the following three steps:

(1) Corpus-assisted in-depth reading.

(2) Students' independent creative writing and continuous writing training.

(3) Corpus-assisted writing feedback.

## *3.4. Case studies of teaching effectiveness*

Before the experiment, in order to know the English level of the students and to determine whether these two classes can be used as subjects. In this paper, the students were tested with the PRETCO--Level A question paper of June 2024 and the corresponding data were collected. The data was first tested for normality through SPSS software and then an independent samples t-test was conducted to determine if there was a large difference in the scores of the two classes.

The pre-test scores of both classes were normally distributed and their significance levels were both > 0.05, there was no significant difference between the scores and they could be used for subsequent experimental studies. The results of the independent samples t-test also found that the Sig values of Levene's chi-square test of students' scores in the two classes were 0.6314 and 0.5928 respectively, both of which were >0.05, indicating that the English proficiency of the students in the two classes was comparable and could be used as the subjects of the teaching experiment.

In addition, this paper also analyzes the differences between the two classes of students in various achievement scores, and also finds that there is not much difference between the two classes in listening, vocabulary and grammar, reading, translation and writing. It indicates that the overall level of the students in the two classes in the applied skills of listening, vocabulary and grammar, reading comprehension, translation and writing is comparable and there is no significant difference, which can be used as subjects.

### 3.4.1. Statistics and analysis of overall performance

After one semester of practicing the corpus-assisted method, this paper tested the students with the PRETCO--Level B test paper of June 2025 and collected the corresponding experimental data. After that, this paper conducted a normality test on the data using SPSS software, and then an independent samples t-test was conducted to determine whether the results of these two classes showed significant differences, in order to understand the effect of corpus-assisted method in improving the English performance of non-English majors in colleges and universities.

The statistical results of the post-test scores of the experimental and control classes are shown in Table 3. It can be seen that the level of significance is higher than 0.05 for both experimental and control classes by Kolmogorov-Smirnov and Shapiro-Wilk tests. And the histograms of the grades of both the classes also show a more normal distribution on the whole. Thus, it can be found that the grades of both classes are normally distributed and the grades can be used as parametric tests.

***Table 3.*** Post-test scores of experimental and control classes

| Object to test | Test specification | Experimental class post-test scores | Control class post-test scores |
|---|---|---|---|
| Kolmogorov-Smirnov[a] | Statistics | 0.1071 | 0.0835 |
| | df | 53 | 50 |
| | Sig. | 0.1902 | 0.1917 |
| Shapiro-Wilk | Statistics | 0.9746 | 0.9756 |
| | df | 53 | 50 |
| | Sig. | 0.1262 | 0.288 |

The statistical results of the posttest scores of the experimental and control classes are shown in Table 4. The mean of the posttest scores of the experimental class was 74.0534, while that of the control class was 66.1892, which was 7.8642 points higher, indicating a more significant increase in the overall English proficiency of the students in the experimental class through the corpus-assisted method. The post-test means of both classes increased, with the means of the experimental and control classes (pre-test means of 61.3724 and 62.7955, respectively) increasing by 12.681 and 3.3937 points, respectively, which illustrates that the experimental class improved significantly more than the students in the control class.

***Table 4.*** The statistical results of the post-test scores of the two classes

| Post-test scores | Experimental class | Control class |
|---|---|---|
| N | 53 | 50 |
| Mean | 74.0534 | 66.1892 |
| Standard deviation | 11.4261 | 12.5375 |
| Standard error of the mean | 1.6985 | 1.9073 |

The results of the independent samples t-test for the posttest scores of the experimental and control classes are shown in Table 5. The results show that the Sig value of Levene's test of variance chi-square is 0.3164 > 0.05, indicating that the variance of the posttest scores of the two classes is the same on the variable of achievement, which meets the conditions of the assumption of variance chi-square for parametric tests. According to the t-test results of the mean equation, the probability of significance is 0.0081 < 0.05, indicating that there is a significant difference between the experimental class and the control class in terms of post-test scores. The difference in the overall English level of the two classes is obvious, which shows that after one semester of corpus-assisted English teaching practice in colleges and universities, the performance of the experimental class is better than that of the control class, and the progress is greater, and Hypothesis 1 proposed in this paper is valid, i.e., "Corpus-assisted teaching can improve the English language proficiency of students in colleges and universities".

***Table 5.*** Independent samples t-test results for post-test scores of two classes

| Test method | Test specification | | Post-test: Assume equal variance | Result: Assume unequal variances |
|---|---|---|---|---|
| Levene's test for variance | F | | 1.0158 | - |
| | Sig. | | 0.3164 | - |
| T-test for the mean equation | t | | 2.5954 | 2.5843 |
| | df | | 102 | 100 |
| | Sig. (both sides) | | 0.0081 | 0.0081 |
| | Mean Difference | | 6.4951 | 6.4951 |
| | Standard error | | 2.2054 | 2.2033 |
| | 95% confidence interval for the difference | Lower limit | 1.2963 | 1.3056 |
| | | Superior limit | 10.7998 | 10.8005 |

### 3.4.2. Statistics and analysis of the results of each question type

In order to further understand the differences in the corpus-assisted approach in improving students' English language application skills in each item, this paper also statistically and analytically analyzes the various scores of the post-test scores of the experimental and control classes to understand the different effects of the corpus-assisted approach on improving students' English language application skills in the items of Listening, Vocabulary and Grammatical Structures, Reading, Translating, and Writing.

The statistical results of the post-test scores of the experimental and control classes for each question type are shown in Table 6. It can be seen that in listening, vocabulary and grammar, reading, translation and writing, the mean differences in the posttest scores between the experimental and control classes are 0.548, 3.9904, 1.205, 1.789 and 2.737 points, respectively. The difference was smaller in listening, largest in vocabulary and grammar, and ranged from 1.205 to 2.737 points in the other question scores.

***Table 6.*** Statistical results of post-test scores for each question type in two classes

| Question types | Respondent | N | Mean | Standard error | Standard error of the mean |
|---|---|---|---|---|---|
| Hearing | Experimental class | 53 | 15.942 | 3.8368 | 0.4883 |
| | Control Class | 51 | 15.394 | 3.5855 | 0.5292 |
| Vocabulary and Grammar | Experimental class | 53 | 10.441 | 2.3232 | 0.3529 |
| | Control Class | 51 | 6.4506 | 2.8097 | 0.3341 |
| Read | Experimental class | 53 | 24.987 | 5.0652 | 0.6925 |
| | Control Class | 51 | 23.782 | 5.359 | 0.7531 |
| Translate | Experimental class | 53 | 10.693 | 2.4313 | 0.3464 |
| | Control Class | 51 | 8.904 | 2.5273 | 0.4164 |
| Writing | Experimental class | 53 | 7.998 | 2.7067 | 0.3782 |
| | Control Class | 51 | 5.261 | 3.0171 | 0.4067 |

The results of the independent samples t-test for each of the post-test scores are shown in Table 7. The significance of the difference between the scores of the experimental and control classes in each of the English questions was not the same.The Sig values of the Levene's chi-square test for each of the questions were 0.7146, 0.205, 0.4726, 0.2618, and 0.7491>0.05, which reflected that the variances of the achievement variables were equal in these five questions, and that the data as a whole were normally distributed. However, the t-test of the mean equations of the scores showed that there was no significant difference in the scores on the listening and reading questions, with P-values of 0.6017 and 0.2135 > 0.05, respectively, while on vocabulary and grammatical structures, translation, and writing, there was a significant difference in the scores, with P-values of 0.0001, 0.0374, and 0.0172 < 0.05, respectively. After one semester of experimenting with the corpus-assisted method, the overall English proficiency of the students in the experimental class has been significantly improved, especially in is significant effect in improving students' vocabulary and grammatical structures, translation and writing. It can be seen that students' English-Chinese translation level can be improved by utilizing corpus-assisted college students on this topic.

***Table 7.*** Independent samples t-test results for post-test performance metrics

| Question types | Assumed variance | Levene's test for variance | | T-test for the mean equation | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F | Sig. | t | df | Sig. (both sides) | Mean Difference | Standard error | 95% confidence interval for the difference | |
| | | | | | | | | | Lower limit | Superior limit |
| Hearing | Equation | 0.05 | 0.7146 | 0.416 | 104 | 0.6017 | 0.303 | 0.727 | -1.128 | 1.696 |
| | Inequality | | | 0.411 | 104 | 0.6017 | 0.329 | 0.706 | -1.118 | 1.717 |
| Vocabulary and Grammar | Equation | 1.01 | 0.205 | 5.348 | 104 | 0.0001 | 2.715 | 0.517 | 1.706 | 3.696 |
| | Inequality | | | 5.362 | 104 | 0.0001 | 2.707 | 0.506 | 1.703 | 3.689 |
| Read | Equation | 0.27 | 0.4726 | 1.219 | 104 | 0.2135 | 1.237 | 1.032 | -0.75 | 3.236 |
| | Inequality | | | 1.226 | 104 | 0.2135 | 1.244 | 1.013 | -0.756 | 3.293 |
| Translate | Equation | 1.29 | 0.2618 | 2.076 | 104 | 0.0374 | 1.161 | 0.56 | 0.052 | 2.257 |
| | Inequality | | | 2.102 | 104 | 0.0374 | 1.139 | 0.56 | 0.049 | 2.236 |
| Writing | Equation | 0.047 | 0.7491 | 2.344 | 104 | 0.0172 | 1.307 | 0.574 | 0.232 | 2.462 |
| | Inequality | | | 2.338 | 104 | 0.0172 | 1.338 | 0.564 | 0.2 | 2.474 |

## 4. Conclusion

In order to improve the quality of English teaching in colleges and universities and the English

translation level of students, this paper first constructs a Chinese-English parallel corpus, and then adopts the Spark-based plain Bayesian text categorization parallelization algorithm to categorize the text of the corpus; on the basis of which, it designs a corpus-assisted college and university English teaching model, and conducts a case study to test the application effect of the model. The results show that:

The algorithm in this paper has good classification effect on the experimental dataset, and the model is stable, the deviation of the results between different classes is small, and its application performance in parallel text classification algorithm is good. In addition, the application of corpus-assisted college English teaching model in teaching can significantly improve students' overall performance, and significantly improve students' English application ability in the three categories of "vocabulary and grammatical structure, translation and writing". It is of great significance to improve the quality of English teaching in colleges and universities and students' English application level.

The focus of this study is to investigate the effects of the corpus-assisted reading and writing teaching model on the reading and writing performance, interest and independent learning ability of the students in the class as a whole. If the study is further refined to study the effects of this teaching mode on students with different English bases, it can not only improve this teaching mode and make the class more efficient, but also provide more personalized teaching to students with different bases.

**References**


1. Nakhli, H. (2021). Developing students' translation competence: The role of tasks and teaching activities. International Journal of Linguistics, Literature and Translation, 4(11), 119-128.
2. Ismail, H., Syahruzah, J. K., & Basuki, B. (2017). Improving the students' reading skill through translation method. Journal of English Education, 2(2), 124-131.
3. Karimian, Z., & Talebinejad, M. R. (2013). Students. Use of Translation as a Learning Strategy in EFL Classroom. Journal of Language Teaching and Research, 4(3), 605.
4. Purwanto, A., Suseno, M., & Setiadi, S. (2023). Integrating Basic Translation Skills and 21st Century Skills in Translation Course. Prosiding Konferensi Berbahasa Indonesia Universitas Indraprasta PGRI, 194-210.
5. Sinambela, E., Siregar, R., & Pakpahan, C. (2023). Improving students' ability in using English with a simple translation: A case on elementary school level. Jurnal Obsesi: Jurnal Pendidikan Anak Usia Dini, 7(3), 3267-3278.
6. Zhao, N., Gao, F., & Yang, D. (2018). Examining student learning and perceptions in social annotation-based translation activities. Interactive Learning Environments, 26(7), 958-969.
7. Bin-Hady, W. R. A., Al-Ahdal, A. A. M. H., & Abdullah, S. K. (2024). The effect of pretranslation techniques in developing EFL students' translation ability. Journal of Applied Research in Higher Education, 16(4), 1176-1187.
8. Priya, T. A., & Jayasridevi, B. (2018). Integrating translation in classroom: Facilitating language skills. Rupkatha Journal on Interdisciplinary Studies in Humanities, 10(1), 118-127.
9. Neves, J. (2022). Project-based learning for the development of social transformative competence in socially engaged translators. The Interpreter and Translator Trainer, 16(4), 465-483.
10. Viorela-Valentina, D. I. M. A. (2021). Translation practice–A means for enhancing student employability. Dialogos, 22(38), 215.
11. Yang, P. (2015). Enhancing intercultural communication and understanding: Team translation project as a student engagement learning approach. International Education Studies, 8(8), 67-80.
12. Yahong, L. (2021). Several Problems and Their Solutions in Translation Teaching for English Majors in Colleges and Universities. Journal of Frontiers of Society, Science and Technology, 1(3), 126-128.
13. Al-Jarf, R. (2022). Positive psychology in the foreign language and translation classroom. Journal of Psychology and Behavior Studies (JPBS), 2(1), 50-62.
14. Yao, Q. (2017, December). China's College English Translation Teaching: Importance, Problems and Suggestions. In 2017 2nd International Conference on Education, Management Science and Economics (ICEMSE 2017) (pp. 217-219). Atlantis Press.
15. Dai, H., & Chen, Z. (2016, August). The Countermeasures and Existing Problems of Translation Teaching in College English. In 2016 International Conference on Humanity, Education and Social Science (pp. 352-355). Atlantis Press.
16. Zheng, Y. (2021). Strategies to improve the effectiveness of college English translation teaching. Advances in Vocational and Technical Education, 3(2), 86-91.
17. Sun, Q. (2021). Common problems in translation practice of english majors and their enlightenment to teaching. International Journal of Social Sciences in Universities, 4, 252-255.
18. Zainudin, I. S., & Awal, N. M. (2012). Translation techniques: Problems and solutions. Procedia-Social and Behavioral Sciences, 59, 328-334.
19. Siregar, R. (2018). Exploring the Undergraduate Students Perception on Translation--A Preliminary Step to Teach Translation in EFL Classes. English Language Teaching, 11(9), 90-101.
20. Al-Ahdal, A. A. M. H., Alfallaj, F. S. S., Al-Awaid, S. A. A., & Al-Mashaqba, N. J. A. H. (2017). Translation courses at Qassim University, Saudi Arabia: A study of existing problems and possible solutions. US-China Foreign Language, 15(3), 45-53.

21. Ganjalikhanizadeh, M., & Fatehi Rad, N. (2022). Technology in Teaching Translation: Problems and Challenges of Current State of Teaching Translation in Post-graduate Studies. International Journal of Foreign Language Teaching and Research, 10(42), 105-118.
22. Tran, P., Nguyen, T., Vu, D. H., Tran, H. A., & Vo, B. (2022). A method of Chinese-Vietnamese bilingual corpus construction for machine translation. IEEE Access, 10, 78928-78938.
23. Zhang, F. (2021). Application of data storage and information search in english translation corpus. Wireless Networks, 1-11.
24. Wang, X. (2021). Building a parallel corpus for English translation teaching based on computer-aided translation software. Computer-Aided Design & Applications, 18.
25. Pan, B., & Qin, Q. (2022). Construction of parallel corpus for english translation teaching based on computer aided translation software. Computer-Aided Design & Applications, 19.
26. Fois, E. (2023). Redefining English language teaching in translator training through corpus-based tasks. Instrumentalising Foreign Language Pedagogy in Translator and Interpreter Training: Methods, goals and perspectives, 161, 112.
27. Alotaibi, H. M. (2017). Arabic-English parallel corpus: A new resource for translation training and language teaching. Arab World English Journal (AWEJ) Volume, 8.
28. Yu, R. (2020). Application of parallel corpus in translation teaching. In 2nd International Education Technology and Research Conference (IETRC 2020) (pp. 736-741).
29. Romli, T. R. M., & Jumingan, M. F. (2015). Open Source Corpus as a Tool for Translation Training. EUROPEAN CENTER FOR SCIENCE EDUCATION AND RESEARCH, 11, 191.
30. Frérot, C. (2013). Incorporating translation technology in the classroom: Some benefits and issues on using corpora and corpus-based translation tools. In Tracks and treks in translation studies (pp. 143-166). John Benjamins Publishing Company.
31. Granger, S., & Lefer, M. A. (2018). MUST: A collaborative corpus collection initiative for translation teaching and research. CECL Papers, 72-73.
32. Fois, E. (2021). Translator training, English language teaching and corpora: Scenarios and applications. International Journal of Language Studies, 15(4), 59-78.
33. Buendía-Castro, M., & López-Rodríguez, C. I. (2013). The web for corpus and the web as corpus in translator training. New voices in translation studies, 10(1), 54-71.
34. Rodríguez-Inés, P., & Gallego-Hernández, D. (2016). Corpus Use and Learning to Translate, almost 20 years on. Cadernos de Tradução, 36(spe), 09-13.
35. Hassani, G. (2011). A corpus-based evaluation approach to translation improvement. Meta, 56(2), 351-373.
36. Alfuraih, R. F. (2020). The undergraduate learner translator corpus: a new resource for translation studies and computational linguistics. Language Resources and Evaluation, 54(3), 801-830.
37. Alhassan, A., Sabtan, Y., & Ismail Omar, L. (2021). Using parallel corpora in the translation classroom: moving towards a corpus-driven pedagogy for Omani translation major students. Arab World English Journal (AWEJ) Volume, 12.
38. Frérot, C. (2016). Corpora and corpus technology for translation purposes in professional and academic environments. Major achievements and new perspectives. Cadernos de Tradução, 36, 36-61.
39. Talhakul, W. (2015). A Corpus-based Approach to Teaching Translation: Can it be implemented in Thai Undergraduate Classrooms?. NIDA Journal of Language and Communication, 20(24), 63-85.
40. Awal, N. M., Ho-Abdullah, I., & Zainudin, I. S. (2014). Parallel corpus as a tool in teaching translation: Translating English phrasal verbs into Malay. Procedia-Social and Behavioral Sciences, 112, 882-887.
41. Poirier, É. (2016). Exploring theoretical functions of corpus data in teaching translation. Cadernos de traduçao, 36(spe), 177-212.
42. Moratto, R. (2023). Corpus-assisted translation teaching: Issues and challenges. Translation & Interpreting, 15(1), 293-297.
43. Gong, Y., & Cheng, L. (2023). Research on the application of translation parallel corpus in interpretation teaching. ACM Transactions on Asian and Low-Resource Language Information Processing.
44. David Gauhl, Kevin Kakkanattu, Melbin Mukkattu & Thomas Hanne. (2025). Integrating Large Language Models with near Real-Time Web Crawling for Enhanced Job Recommendation Systems. Computers,14(9),387-387. https://doi.org/10.3390/COMPUTERS14090387.
45. Shixiao Li. (2025). Employment trend prediction and innovative talent training strategy optimization using naive Bayes classifier. Journal of Computational Methods in Sciences and Engineering,25(5),3973-3985. https://doi.org/10.1177/14727978251337979.
46. S. Suganya & S. Selvamuthukumaran. (2023). Stacked multi-layer security for Hadoop distributed file system using HSCT steganography. Concurrency and Computation: Practice and Experience,35(21), https://doi.org/10.1002/CPE.7711.