International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 17 (2025) pp. 1-18

https://doi.org/10.70917/ijcisim-2025-0302
Article

A Study on Vocabulary Memory Strategies Based on Cluster Analysis in Business English Vocabulary Learning

Xuefeng Li *

School of Languages and Media, Anhui University of Finance and Economics, Bengbu, Anhui, 233030, China; lxfblcu@126.com

Abstract: The traditional business English vocabulary learning and memorization strategy based on textbook units or vocabulary lists is inefficient, and learners face the problems of heavy memorization burden and easy confusion. This study proposes a vocabulary memorization method based on clustering strategy. The method systematically organizes the vocabulary memorization strategies of phonology, context, word formation, word meaning and clustering. The similarity calculation method and the K-mean algorithm in dividing clusters are used to re-categorize the vocabulary based on the semantics of Business English vocabulary to form vocabulary clusters with a high degree of relevance. Finally, 90 first-year students of the class of 2019 within a university in province A were selected as the research subjects and the vocabulary memorization strategy was applied to conduct a teaching experiment. The experimental data show that the clustering effect of this paper's method is better and can generate more regular vocabulary clusters, and the students in the experimental group who accept the teaching method of thematic clustering presentation are significantly different from their counterparts in the control group in the tests of vocabulary immediate posttest scores, delayed vocabulary scores, etc., and the difference in the mean value between the experimental group's immediate posttest scores and delayed posttest scores is reduced compared with that of the control class by 1.37.This illustrates that the clustering analysis can effectively optimize the memory storage of Business English vocabulary and provides an empirical method for the innovation of vocabulary teaching.

Keywords: Business English vocabulary; K-means algorithm; similarity calculation; vocabulary memorization strategy

1. Introduction

The study and mastery of Business English vocabulary is the basic guarantee for Business English majors to improve their language communication ability and practical application ability. Only by mastering a certain amount of vocabulary and being able to skillfully use the specialized vocabulary in the field of business, can they use English to successfully complete business communication [1-4]. Business English vocabulary belongs to the category of professional vocabulary, in which there are many professional terms and industry terms, and some ordinary words also have special meanings in the business field [5-7]. Therefore, learners of Business English should pay attention to the accumulation of this kind of specialized vocabulary and the application of all kinds of vocabulary in actual business scenes [8-9]. In addition, there are a large number of Business English vocabulary, so it is necessary to pay attention to the skills and methods to master both the meaning and the usage of the vocabulary, and try to expand the amount of positive vocabulary, because the demand for positive vocabulary is much greater than negative vocabulary in the actual business field, which is also a feature of the strong practicability of Business English [10-13].

When learning business English vocabulary, adopting appropriate methods to memorize can get twice the result with half the effort. Traditional vocabulary memorization strategies include categorization, association, practice, etc., and with the development and needs of the industry, the application of cluster analysis further improves the efficiency of vocabulary memorization [14-17].

Cluster analysis is a data mining technique based on pattern recognition and statistical theory, which is used to uncover the so-called “patterns” and knowledge hidden behind the data by letting the items in the data set be grouped into different clusters in a connected way to present their characteristics [18-21]. Cluster analysis is mainly used in the fields of qualitative analysis, pattern recognition, decision analysis, image processing, automatic reasoning, etc., and its main nature belongs to unsupervised learning [22-23]. In Business English vocabulary learning, clustering analysis is able to automatically group vocabulary according to word meanings, usage scenarios and other characteristics through algorithms, which makes the similarity within the group high and the difference between the groups high, i.e., the correlation between Business English vocabulary is utilized to improve the effect of vocabulary memorization, which is of great significance for improving the efficiency of Business English vocabulary learning [24-28]. Literature [29] examined the application of using semantic and clustering methods in English classrooms aiming to improve students' vocabulary, and revealed the effectiveness of the above two methods in vocabulary teaching through experiments. Literature [30] examined the effectiveness of cluster analysis in improving students' vocabulary acquisition and based on experiments showed that cluster analysis positively affected writing skills by improving students' vocabulary learning.

However, whether in business or general English vocabulary learning, cluster analysis is only one of all the methods to improve vocabulary learning; it is not completely applicable to many learners, and only finding a suitable method is an effective strategy to improve English vocabulary memorization. For example, literature [31] examined students' self-strategies in English vocabulary memorization, and the study based on a qualitative survey showed that the strategies used by students in memorizing English vocabulary were strategies such as taking notes and repetition. Literature [32] aimed to find out students' views on the use of memory strategy training to promote vocabulary memorization, and a survey of the students showed that they believed that the use of mnemonic techniques as a vocabulary learning strategy helped to improve vocabulary memorization, and that this method was more effective than the use of word lists. Literature [33] analyzes the basic features of Business English vocabulary and the impact of corpora on Business English vocabulary teaching, aiming to achieve further development of the subject of Business English with the support of corpora. Literature [34] utilized a questionnaire to examine the vocabulary learning strategies used by students in English classrooms, and the results showed that the deterministic strategy was a common learning strategy used by students, and that there was a positive relationship between this strategy and students' academic performance. Literature [35] highlighted the failure of some students to achieve a passing level in English language tests and these students summarized the reasons as limited time to prepare for the test and insufficient English vocabulary. Literature [36] discusses strategies regarding English vocabulary learning and based on the literature review states that learners' vocabulary was significantly improved through the use of appropriate strategies.

In addition, literature [37] describes the impact of Memrise, an online learning platform, on students' learning of English vocabulary, based on a study of a business English course for management majors that showed its positive effect in facilitating students' learning of new vocabulary and its meanings, among other things. Literature [38] aimed to understand students' vocabulary learning strategies based on a qualitative research method revealed that students' vocabulary learning strategies include monolingual and bilingual dictionary use, language media, and daily conversation. Literature [39] investigated the vocabulary learning strategies commonly used by students of English for Specialized Purposes (ESP), and the analysis of questionnaires and interviews pointed out that students preferred metacognitive strategies and had a positive attitude towards technology as an aid to vocabulary acquisition. Literature [40] emphasizes the importance of foreign language vocabulary learning and the difficulties in vocabulary memorization and describes the effectiveness of mobile-assisted language learning in improving vocabulary memorization. Literature [41] emphasizes the importance of having a broad English vocabulary, especially high-frequency vocabulary, and points out that in order to achieve this, English learners are constantly exploring effective vocabulary learning strategies. Literature [42] analyzes the English vocabulary learning strategy of learning cards, emphasizes the effectiveness of this method in improving students' vocabulary memory, classroom participation and motivation, and promotes the use of card teaching strategy English vocabulary memory courses.

Current business English teaching and learning memory strategies are mostly stuck in the traditional mode of mechanical repetition and discrete memorization, which makes it difficult for learners to systematically form a vocabulary knowledge system. In order to overcome the above limitations, this study introduces the cluster analysis method and constructs a complete research framework from strategy sorting, algorithm clustering to teaching verification. Firstly, the similarity calculation method and the K-mean clustering algorithm are applied to the reorganization of Business English vocabulary, and regular clustered vocabulary clusters are formed through thematic associations between words. Finally, a

rigorous instructional controlled experiment was designed to verify the impact of the strategy on enhancing learners' vocabulary acquisition effect and memory quality.

2. Business English Vocabulary Memorization Strategies

2.1. Speech strategy

Phonics strategy is an efficient vocabulary learning and memorization strategy that utilizes the phonetic symbols of vocabulary and the laws of pronunciation. The basis of English vocabulary learning is to master phonetics, and correctly spelled phonetics is the prerequisite for correct vocabulary spelling, and the degree of correct phonological spelling is usually directly proportional to the degree of vocabulary mastery. Phonetics is one of the main reasons for the bipolar gap between high school students in English learning. The problem lies in the vocabulary learning process where students' phonological spelling skills are not up to scratch, and they can't memorize words with the same phonological and morphological memorization rules, and they usually memorize vocabulary based on rote memorization of phonetic pronunciation or the Chinese style of labeling of pinyin of Chinese characters, so that the spelling of the words loses its accurate and reliable basis and guarantee, and the learning of vocabulary becomes a time-consuming and inefficient part of English language learning. Vocabulary learning becomes a time-consuming and inefficient part of English learning.

English is a pinyin language, and business English teaching requires students to systematically master basic phonetic knowledge, such as the pronunciation of vowel letter combinations in stressed and light syllables, the pronunciation of consonant affixes, etc., and the changes in pronunciation, such as legato, bursting, and weak pronunciation, etc. And based on this, students are required to acquire the basic phonetic knowledge. And based on this, students are required to acquire the relationship between phonetics and spelling, so that students learn to acquire vocabulary based on the phonological relationship through phonics strategies, and improve the accuracy and quantity ratio of memorized vocabulary. Phonics strategy is based on the relationship between pronunciation and spelling, comparing the same or similar points of vocabulary in terms of sound, shape and meaning and finding out the pattern, thus reducing the difficulty of learning words to a certain extent and improving the memorization effect.

The phonological strategy is in line with the teaching requirements of “regression”, that is, English words are not isolated, vocabulary learning needs to reflect on, and has the characteristics of openness, double-sided and interpretive; at the same time, the strategy also reflects the teaching requirements of “relevance”, focusing on the use of The strategy also reflects the “relevance” of the teaching requirement, focusing on the utilization of the intrinsic correlation between vocabulary and phonology in vocabulary acquisition.

2.2. Situational strategies

The context of teaching refers to the emotional atmosphere created by the teacher in teaching, vocabulary learning is a certain degree of teaching activities with an emotional atmosphere, a good context can fully mobilize students to learn vocabulary, inspire their learning thinking, expand their learning limitations, and is an important strategy to improve the effectiveness of vocabulary teaching in high school. In the process of vocabulary learning, a specific situation is created in advance, such as “at the airport”, and then the vocabulary related to the situation is associated with the mind as shown in Figure 1.

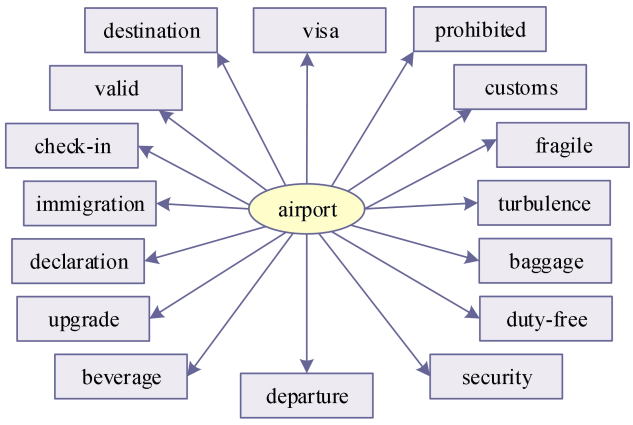


Figure 1. Vocabulary related to the situation in the airport.

The creation of context can also be done with the help of intuitive videos, images or objects. Appropriate and imaginative means of creating context can help to enliven the atmosphere, stimulate students' thinking, reduce the immediate pressure of learning vocabulary, and eliminate the boredom of memorizing vocabulary. It can inspire students to think about vocabulary learning, guide them to go deeper into the vocabulary memorization situation, and make them actively and smoothly engage in vocabulary acquisition strategies in line with “richness”, with depth and meaning of multi-level, and a certain degree of dispersion. At the same time, the contextualization strategy also embodies “relevance”, reflecting the intrinsic connection in vocabulary learning and the cultural connection beyond vocabulary, which are complementary to each other.

2.3. Word formation strategies

In terms of lexicography, English words belong to morphological structure. Morphemes are divided into real morphemes and grammatical morphemes, real morphemes are the semantic basis of the word, and some of them can be used as words alone, such as the root of the word belongs to the real morphemes; while grammatical morphemes are only used to express the meaning of the additional meaning and grammatical significance of the need to combine with other morphemes in order to form words, such as affixes are grammatical morphemes. Suffixes can be divided into prefixes and suffixes, and the change of prefixes generally involves only a change in the meaning of the word; the change of suffixes will change the meaning of the word, and at the same time will change the nature of the word. Therefore, business English vocabulary teaching should focus on the knowledge of word formation and the meaning of high-frequency affixes, so that students can form the regularity and systematization of the memory method in vocabulary learning, so as to learn by example, and carry out batch vocabulary acquisition in a point-by-point manner, and ultimately realize the purpose of expanding the vocabulary effectively.

2.4. Lexical strategies

Lexical strategies focus on vocabulary acquisition that extends from the original meaning of a word to its expanded meaning. English vocabulary often has multiple meanings, so taking its original or basic meaning as a starting point and guiding students to speculate on its derivational and metaphorical meanings can lead to a deeper understanding and accelerated mastery of the vocabulary.

2.5. Clustering Strategy

In high school English vocabulary learning, vocabulary can be grouped into clusters according to topics or themes in order to consolidate the internalization of learned vocabulary and accelerate the internalization process. This strategy is in line with the “relevance”, using some kind of correlation between vocabulary for diversified combinations of clustering, and realizing the effect of learning vocabulary by analogy or learning by example. In conclusion, English vocabulary learning is regular, and the emphasis on openness, self-organization, interactivity and process is highly applicable to high school English vocabulary teaching, and has a strong guiding role in the exploration of vocabulary learning strategies.

3. Study design

3.1. Research questions

Vocabulary is the basis for mastering a language and vocabulary learning is an important part of second or foreign language learning. Learners and teachers spend a lot of time and energy on vocabulary learning and vocabulary teaching. As scholars at home and abroad continue to deepen their research on vocabulary teaching, the issue of vocabulary presentation has attracted more attention, and the academic community has not yet come to a conclusion as to what kind of presentation is more acceptable to learners. The main focus of the debate is on the advantages and disadvantages of various vocabulary presentation methods and whether their advantages and disadvantages are conditionally limited. The purpose of this study is to find the most effective way of presenting English vocabulary for English teachers by comparing the effects of clustered vocabulary presentation (thematic clustering and semantic clustering) and traditional vocabulary (vocabulary lists) presentation on students' English vocabulary learning. On this basis, the research questions of this paper are formulated as.

(1) Is there a significant difference between the vocabulary memorization effect of the experimental class and that of the control class after teaching vocabulary through clustered vocabulary presentation?

(2) Which presentation method performs better in immediate testing and delayed retention in terms of vocabulary acquisition between the clustered vocabulary presentation method and the traditional vocabulary presentation method?

3.2. Subjects of the study

This study was conducted within a university in Province A. The experimental subjects were 90 freshman students of the class of 2019. The experimental subjects were from two classes and the student population was all general high school graduates. Before the experiment, a pre-test of English proficiency is conducted for the students of the two classes, after which the data are counted and analyzed to determine that there is no statistically significant difference and that the English learning levels of the students of the two classes are roughly equivalent before the experiment can be conducted. The author chooses some vocabularies in the public foreign language textbook of the class of 2019 in our school for this teaching experiment study. The vocabulary level test is administered to the students of the two classes participating in the experiment as a pre-test for the experiment. The vocabulary level test paper is based on the British National Corpus which determines a list of 15000 word families and the question type is multiple choice. The author chose the first 2000 word families as the vocabulary level test paper for this study, which consists of 40 questions, with one point for correct answers and a total of 40 points.

The author selected 50 words in the third unit of the first book of the New Generation English Course for the business English vocabulary clustering experiment.

3.3. Research methodology

3.3.1. Experimental methods

The purpose of this study is to investigate the differences in the effects of clustering presentation style and traditional vocabulary presentation style in new word learning. The study utilized an experimental method with a 2×3 two-factor within-subjects design, the experiment contained two variables, the independent variable was the vocabulary presentation method with two levels, level 1 was the cluster presentation method and level 2 was the traditional vocabulary list presentation method, and the dependent variable was the vocabulary memorization effect, which was reflected in the vocabulary test scores of the students of the two classes. Formal tests were divided into immediate tests, which were administered in class on the day of vocabulary presentation, and delayed tests, which were administered two weeks after vocabulary presentation.

The tests were administered after the new vocabulary instruction, and each test was administered in the classroom using ten minutes of classroom time by the instructor, who supervised the test-takers' writing of the test questions and then collected the test papers in the classroom. After scoring, the data were entered into the Statistical Package for the Social Sciences (SPSS) 25.0 to analyze the data, and group statistics and independent samples t-tests [43] were performed on the vocabulary test scores of the experimental and control classes to compare the experimental data of the two groups to see if they were significantly different.

3.3.2. Similarity calculation method

Rationale for similarity calculation [44] in cluster analysis: 1) there are differences between things; 2) all clustering algorithms are based on similarity calculation; 3) after clustering, all things are divided into groups, which makes it easy to identify the characteristics of each group; 4) the characteristics and representations of clusters or groups can be better interpreted; 5) clustering of data can better help in the retrieval of structured information; 6) similarity is the most basic step of clustering , in which the classification of new things mainly relies on similarity computation to complete; 7) Based on the classification information, similarity computation can be applied to predict the behavior of new things; 8) The similarity measure can be used to dig deeper into the structure of the relatively centralized data; 9) Many data mining techniques, such as clustering, classification, and feature selection can be used with similarity computation.

Whether things are similar or not depends on the selection of features. Usually, after the feature selection, the text set is clustered and analyzed, and the measure of text similarity needs to be determined first. For different types of text feature information, the method of similarity calculation will be different. The method usually used is to measure the similarity between the features, and then the features are clustered and analyzed according to the similarity or distance between two and two. Distance measurements are usually used for similarity measurements between two objects. In fact, the choice of distance formula is crucial in cluster analysis.

The distance-based similarity measure mainly includes the following algorithms:

1) Euclidean distance

The most commonly used distance measure between quantitative variables is the Euclidean distance. Usually, ‘distance’ refers to the Euclidean distance, which is the open square of the sum of the squares of the differences between the same features between two things.

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \tag{1}$$

2) Minkowski distance.

$$d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{n} (x_{ik} - x_{jk})^{\lambda}} \tag{2}$$

3) Manhattan distance.

$$d_{ij} = \sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right| \tag{3}$$

In addition to the three distance-based similarity algorithms mentioned above, one of the more commonly used algorithms in language research is the cosine similarity measure.

4) Cosine Clip

Also known as the correlation coefficient, the cosine pinch angle between two vectors is usually measured. Values are often taken between $[-1,+1]$ rather than distance measures. When two vectors are similar, the cosine angle value is high.

$$s_{ij} = \frac{\sum_{k=1}^{n} \left(x_{ik} \times x_{jk} \right)}{\sqrt{\left(\sum_{k=1}^{n} x_{ik}^2 \times \sum_{k=1}^{n} x_{jk}^2 \right)}} \tag{4}$$

The advantages and disadvantages of similarity calculation methods have been discussed by many scholars through empirical studies. According to some scholars, the commonly used geometric algorithms for vector distances include Euclidean distance or Manhattan distance, and five distance measures and cosine similarity computation approaches have been used for lexical similarity for each of the four different tasks, with the best algorithm overall being cosine. Two other common algorithms include logarithmic and Jaccard coefficients. Some scholars believe that Euclidean distance is not suitable for the expression of high-dimensional text similarity due to its extremely poor accuracy as the difference in distance decreases after normalizing the vectors to the unit hypersphere; cosine distance is more suitable for the expression of text similarity because it usually yields a better clustering effect in measuring text similarity. In summary, for different research purposes and research tasks, it is necessary to combine the practical and select the optimal algorithm through continuous attempts.

3.3.3. Delineation of clusters

Due to the simplicity and efficiency of the K mean algorithm [45], it is perhaps the most widely used algorithm for dividing clusters. To date, this classical algorithm has been chosen for many clustering tasks. The core idea of the algorithm is to find K clustering centers $c_1, c_2, \ldots, c_K$ such that the sum of squares of the distances between each data point x_i and its nearest clustering center c_v is minimized.

The clustering process of the K mean method is as follows:

Step 1 Randomly select k objects as the center of the cluster.

Step 2 For the remaining objects, assign a cluster to them (i.e., assign them to the closest cluster) based on their distance from the center of the cluster that exists.

Step 3 Recalculate the mean value for each of the changed clusters.

Step 4 ends if the new mean value is the same as the original mean value, otherwise move to step 2.

Steps 2, 3 of the above process are repeated until the criterion function converges. The criterion function is usually referred to as the residual sum of squares (RSS) and K-means clustering aims to make this function minimized.

Many scholars have summarized the advantages and disadvantages of the K mean clustering algorithm. Its advantages are mainly reflected in the following aspects: ① It requires less space: only data points and centroids need to be stored, with a storage capacity of $O(n+K)$; ② less time requirements, and n linear correlation, namely $O(IKn)$, usually I is very small, and generally $K << n$; ③ It remains efficient even after multiple runs; ④ Simple and applicable to various data

types; ⑤ The output result does not depend on the order of data processing. However, the K mean algorithm also has the following problems: ① It is difficult to select the K value, that is, how many clusters are reasonable to divide it into; ② The selection of the initial centroid has a significant impact on the clustering results. Different initial values converge to different local minima, meaning the algorithm is extremely unstable. ③ It cannot handle non-spherical clusters, clusters of different sizes and densities, and is sensitive to outliers or noise points.

How to choose K value, K mean algorithm aims to choose the optimal K value of the criterion function, that is, make the residual sum of squares (RSS) optimal when the value of K, the smallest value of all the possible output K clustering results of the RSS will be recorded as $RSS_{\min}(K)$, then $RSS_{\min}(K)$ will monotonically decrease as K increases, and $RSS_{\min}(K)$ will take the minimum value 0 when $K = N$, where N is the number of all documents. That is, we should end when every document becomes a cluster. Obviously, this is not the optimal clustering result.

One heuristic approach to the above problem is to evaluate $RSS_{\min}(K)$ using the following method. First we perform i times (e.g., $i = 10$) a clustering process, where the result of each clustering contains K clusters (the initialization is different for each clustering), and then we compute the RSS value of the result of each clustering. We then take the minimum value of i RSS, denoted as $\widehat{RSS}_{\min}(K)$. Now increase K and look for the inflection point in the curve based on the change in the value of $\widehat{RSS}_{\min}(K)$, after which the decrease in $\widehat{RSS}_{\min}(K)$ will be significantly lower.

4. Analysis of business English vocabulary clustering results

4.1. Results of the dissimilarity analysis

K-means algorithm is a classic clustering algorithm, it is more appropriate to choose K-means method to do clustering for such data. Therefore, we choose the K-means method that comes with Matlab to do the cluster analysis and plotting.

The experiment intends to plot three groups of graphs, the following three groups of graphs to do the experiment before the results of the speculation:

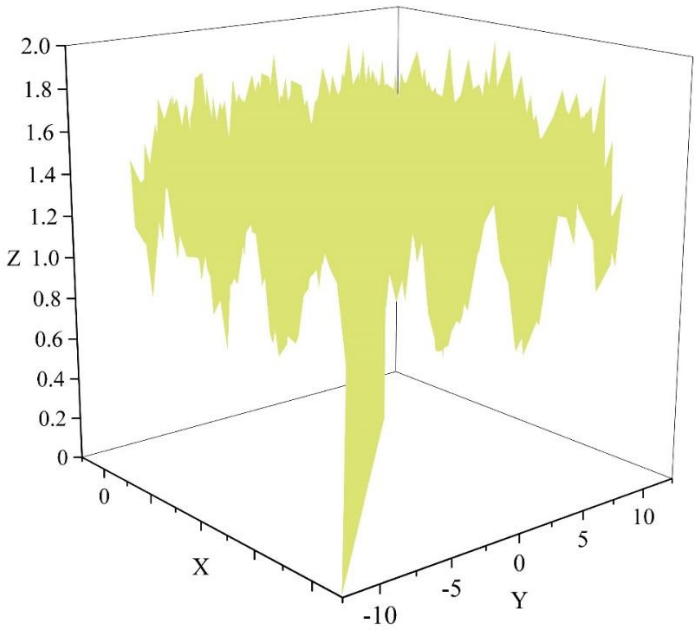
The first group:contains two graphs, one is a three-dimensional graph, with the order of the 50 words before clustering as the x-axis and y-axis, respectively, and the number of times these two words appear at the same time as the b-axis, the three-dimensional graph is plotted. Then conjecture that the value on the diagonal is 0 (specifying that the number of times itself occurs at the same time as itself is 0), and that there should be a staggering distribution of points of different heights on either side of the diagonal. The other is a plan view, with the order of the 50 words before they were clustered as the x-axis and y-axis, respectively, that is, a top view of this three-dimensional map, which can also be said to be an isometric map with the z-axis as a reference. And it can also be seen from these data collected that the contour lines should be mostly concentrated near the origin of the coordinates. Because according to the frequency of occurrence of these words from high to low to do sorting, then the frequency of occurrence of high words, two between the two appear in the same filename probability is also large, of course, should be distributed in the coordinates of the origin of the place close to.

The second group:contains two graphs, one is a three-dimensional graph, with the order of the 50 words after clustering as the x-axis and y-axis, respectively, and the number of times these two words appear at the same time as the z-axis, to plot the three-dimensional graph. Then conjecture that the values on the diagonal are still 0, but after clustering they should appear as clusters of clusters, and these clusters should appear near the diagonal. The other is a plan view, with the order of the 50 words after clustering on the x-axis and y-axis, respectively, which is a top view of this three-dimensional graph, i.e., an isometric graph with the z-axis as the base. After clustering, these clusters should appear around the diagonal.

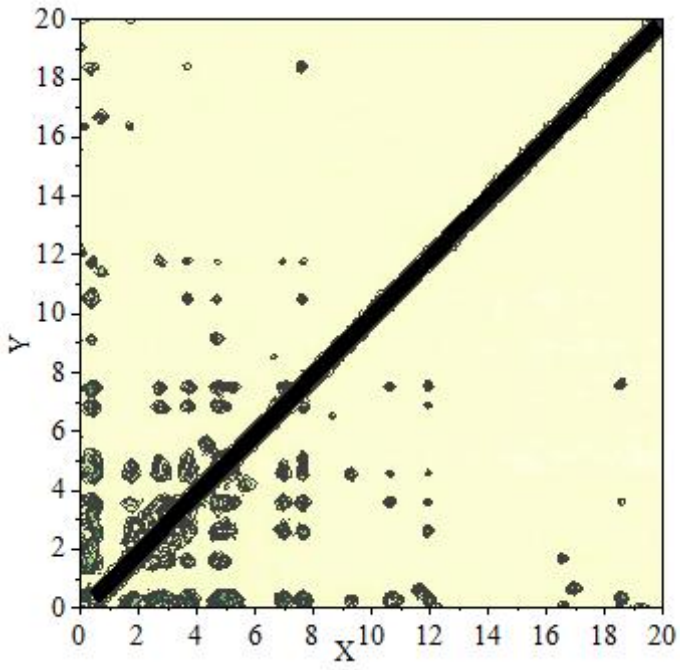
The third group: contains two plots, one that plots the plan view of these 50 words before they are clustered and selects only some of these words that are closer together to depict the points. The other is a plan view of the 50 words after clustering, with only the closer words selected to depict the points. It is conjectured that these representative points after clustering should also appear in the form of clusters and be distributed around the diagonal line, the following plots are drawn by choosing the k-means clustering method with the k value of 5, that is, the words are divided into 5 classes of different types, the first and second sets of plots are shown in Fig. 2 and Fig. 3, respectively.

The (a) figure in Fig. 2 is the stereogram before clustering, and (b) figure is the planar contour map before clustering. From the (b) figure, it can be seen that before unclustering these words are scattered

around the origin of the coordinates with no certain regularity, which is the same as the previous prediction. The (a) figure in Fig. 3 is the stereogram after clustering, and the (b) figure is the planar contour map after clustering. (b) figure has four blue lines horizontally and vertically, which divides the whole coordinate into five regions from left to right, and also divides it into five regions from bottom to top, which represent these five types of words. Then there will be five regions in its diagonal position, which are overlapped horizontally and vertically, and from the previous analysis, it can be seen that these regions should gather all the points after the classification, while the points elsewhere are almost zero. As can be seen from figure (b), these words have shown a tendency to cluster after clustering, and some clusters do cluster around the diagonal, clustering into two clusters in the upper right and lower left corners, respectively.

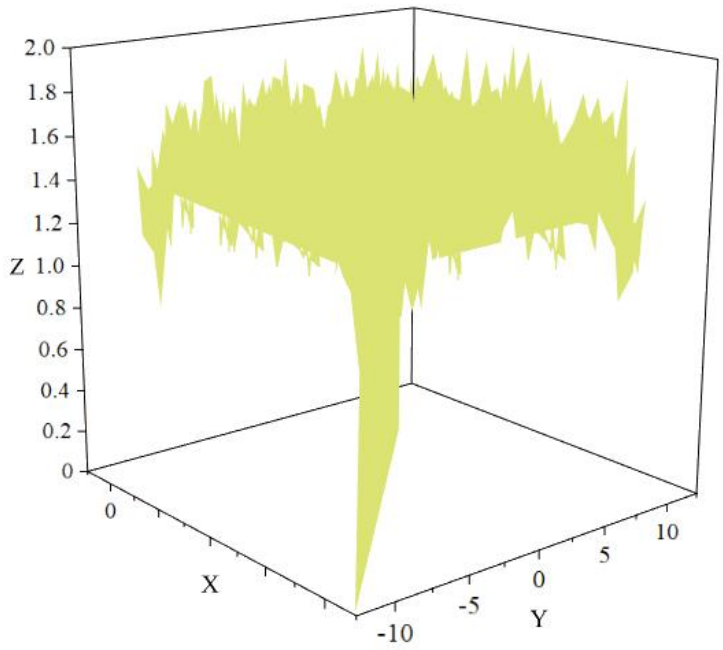


(a)Unclustering and stereogram

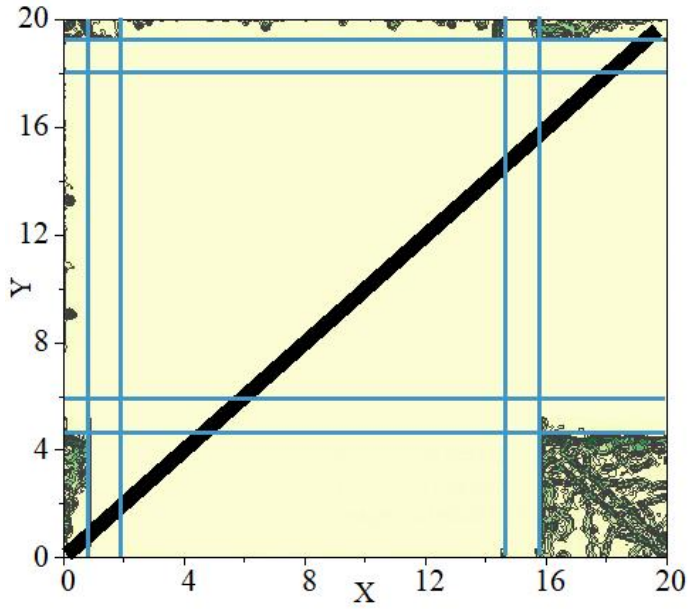


(b)Unclustered floor plan

Figure 2. The flat and stereogram of the unclustering.



(a) Afterclustering and stereogram



(b) The plan after the cluster

Figure 3. After the clustering, the plane and the stereogram.

Figures 4 and 5 show the plan view of certain points before and after clustering, respectively. As can be seen from Figure 4, before clustering these points are scattered within the axes. And from Fig. 5, we can see the effect after clustering, in the upper right corner of the place does have the characteristics of clustering, and clustered in the vicinity of the diagonal. And there are only fewer scattered points away from the diagonal, it can be seen that after eliminating some points with smaller correlation, the clustering effect is more satisfactory, and there are fewer points appearing on both sides away from the diagonal.

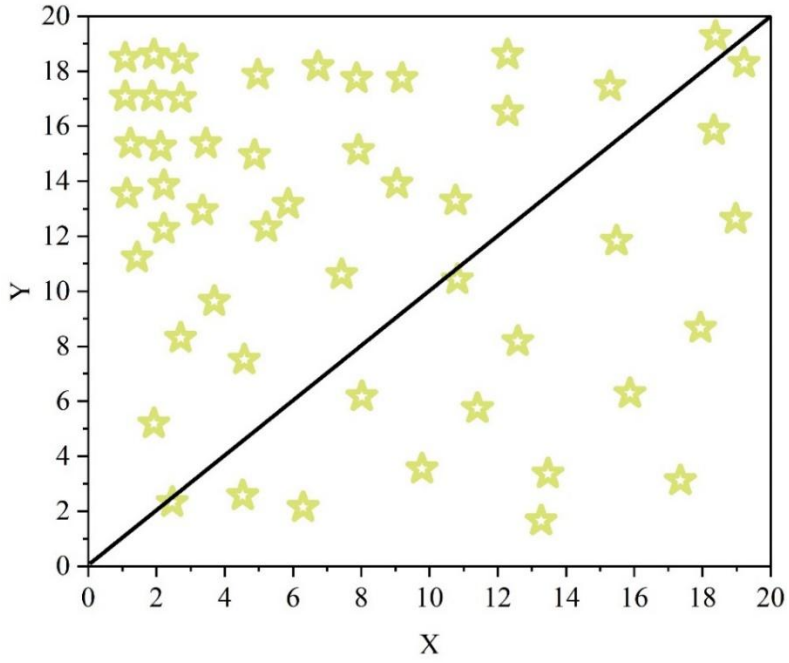


Figure 4. The plane plan of certain dots.

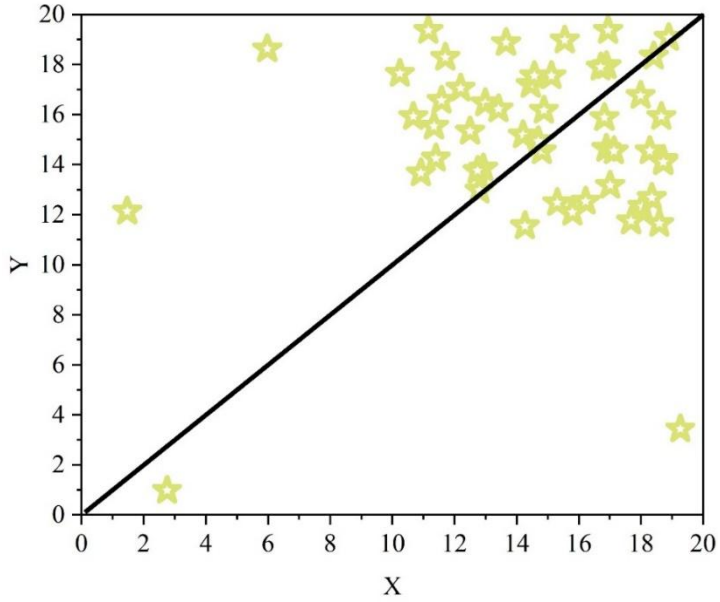


Figure 5. The floor plan after certain points.

4.2. Summary of Cluster Analysis

After analyzing the clustering results of 50 business English words earlier, 800 English words were counted from the database to do the cluster analysis, and Figure 6 shows the distribution frequency of the 800 words. On the one hand, it is to verify the conclusions that have been drawn, and on the other hand, it is to draw a general conclusion. The k-means clustering method is used to do the classification of these 800 words into categories, which are still divided into five categories, however, the results obtained after clustering are not satisfactory, after divided into five categories, only two categories gather more than 90% of the vocabulary, and the other three categories basically do not gather vocabulary. The reason for this is that among these 800 words, only the first 150 words have a high frequency of occurrence, while the frequency of the other words is so low that it can be ignored, which can be seen from the following figure (the horizontal coordinate indicates the 800 words, and the vertical coordinate indicates the

number of times these 800 words appear in the database).

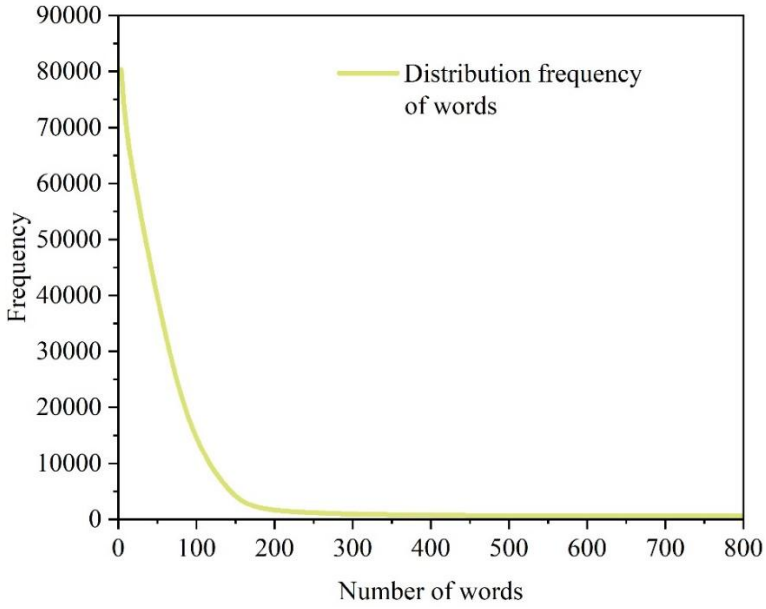


Figure 6. The frequency of the 800 words.

As can be seen from the above figure, the inflection point of this fold line appears at about 150, and the frequency of words after 150 is almost 0, which can be ignored, so the 150 words with the highest frequency of occurrence in the front are taken out from these 800 words to do the clustering analysis, and are still classified into 5 classes by using the k-means clustering method, and the plane results before and after the obtained clustering are shown in Figures 7 and 8. The obtained before and after clustering graphs are shown in Fig. 7 and Fig. 8.

After divided into 5 classes, there is still a very good clustering effect, it is clear to see that the 5 classes of words divided by the blue line, each class shows a different distribution curve, which proves that these words do have certain classification characteristics, and also proves the correctness of the previous experimental conclusions. Since these 150 words are the top 150 words with the highest frequency of occurrence and are unscreened, such a categorization process is more general.

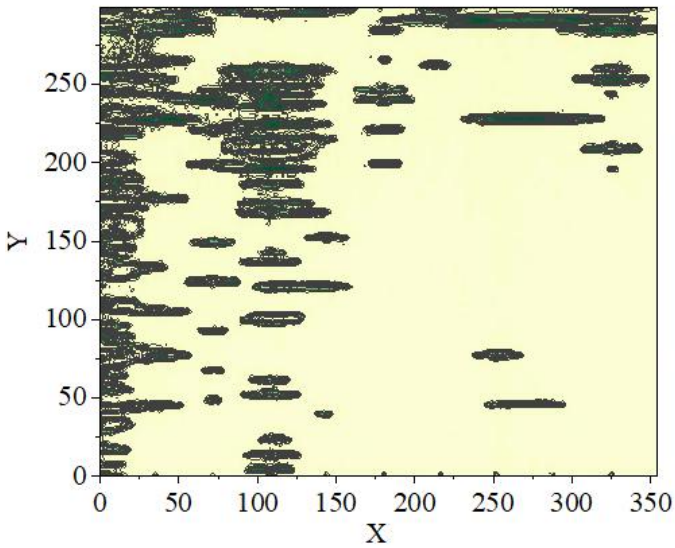


Figure 7. The plane results of the clustering front.

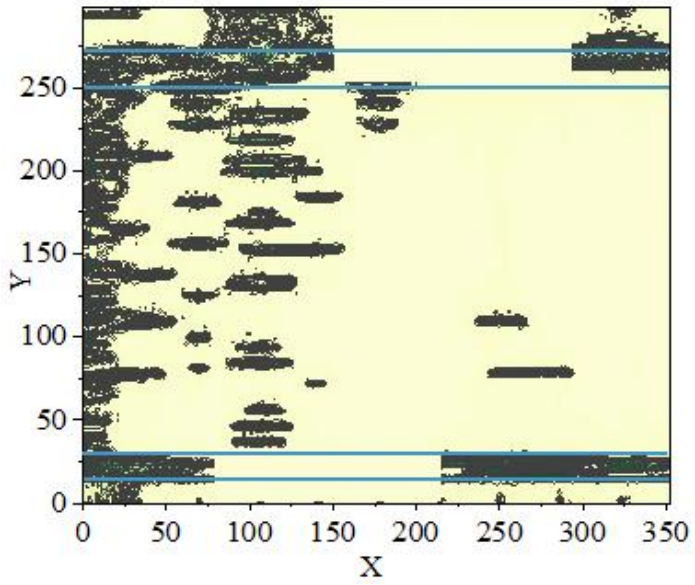


Figure 8. The plane results of the clustering.

5. The effect of business English vocabulary clustering on student achievement

5.1. Effect of Thematic Clustering Presentation on Students' Vocabulary Achievement

The author conducted an independent samples t-test analysis on the business English instant posttest scores of students in two classes to investigate whether the vocabulary presentation of thematic clustering is more effective than glossary presentation in improving students' business English vocabulary scores, and the results are shown in Table 1.

In the Business English Vocabulary Immediate Posttest, the experimental class scored higher than the control class, with a difference of 3.529 between the means of the two groups.The Levene's variance test showed that the probability of significance (Sig) value was 0.705, which was greater than 0.05, indicating that the two classes had equal variance in the variable of vocabulary immediate posttest scores. There is a significant difference between the experimental and control classes on the immediate posttest vocabulary scores. That is, there is a significant difference between the vocabulary scores of the two classes under thematic clustering and vocabulary list presentation, and students' vocabulary learning is significantly better under thematic clustering presentation. As shown in Table 2:

Table 1. Statistics on the results of the vocabulary.

	Class	Number	Mean value	Standard deviation	Standard error mean
Vocabulary achievement	Laboratory class	45	34.785	6.4521	1.0364
	Control class	45	31.256	6.0213	0.9258

Table 2. Test of independent sample t after real-time analysis.

	Variance equivalence test		The average value of t is tested						
	F	Sig.	t	df	Sig. (2-tialed)	Mean Difference	Std.Error Difference	95%CI Lower	95%CI Upper
Assumed	0.144	0.705	1.996	90	0.045	2.7785	1.3895	0.0106	5.5378

Equal variances									
Not Assumed Equal variances			1.996	89.451	0.051	2.7786	1.3945	0.0034	5.4401

5.2. The effect of topic clustering presentation on the learning effect of different vocabularies

The author discusses the effects of glossary presentation and thematic clustering presentation on the learning effect of different lexical words by analyzing the vocabulary scores of Business English immediate posttest, the students' immediate posttest scores of the two classes were classified according to the lexical features and analyzed by the independent samples t-test respectively, and the statistical results of the different vocabulary scores and the results of the independent samples t-test are shown in Tables 3 and 4, respectively.

Noun scores, verb scores and adjective scores of the experimental class were higher than those of the control class, and the difference in the means of noun, verb and adjective scores was 1.633, 1.102 and 0.828, respectively.

In the independent samples t-test, which focuses on the variability of the learning effects of nouns, verbs and adjectives in different presentation styles, the Levene's variance test shows that the probability of significance (Sig) values are 0.285, 0.241 and 0.675, which are greater than 0.05, indicating that the two classes have equal variance in the three variables of nouns, verbs and adjectives, respectively.

Table 3. Statistics of different lexical scores.

Grade	Class	Number	Mean value	Standard deviation	Standard error mean
Noun score	Laboratory class	45	12.089	2.2806	0.3609
	Control class	45	10.456	2.8596	0.4411
Verb grade	Laboratory class	45	10.689	2.1658	0.3456
	Control class	45	9.587	1.8795	0.2897
Adjective achievement	Laboratory class	45	11.023	2.1856	0.3451
	Control class	45	10.195	2.1435	0.3318

Table 4. Independent sample t test for different lexical results.

	Variance equivalence test		The average value of t is tested						
	F	Sig.	t	df	Sig.(2-tialed)	Mean Difference	Std.Error Difference	95%CI Lower	95%CI Upper
Noun score	1.154	0.285	2.456	90	0.015	1.4081	0.5739	0.2687	2.5489

Verb grade	1.356	0.241	2.478	90	0.015	1.1125	0.4476	0.2156	1.9932
Adjective achievement	0.179	0.675	1.693	90	0.097	0.8102	0.4785	-0.1436	1.7655

For words of different lexical properties (nouns, verbs, and adjectives), this study conducted a one-way ANOVA on the data results of the immediate posttest to compare the learning effects of nouns, verbs, and adjectives under the vocabulary presentation of thematic clustering. The results of the effect of thematic clustering presentation on vocabulary learning of different lexical properties are shown in Tables 5 to 7.

Table 5 shows the descriptive statistical analysis of the target vocabulary under the thematic clustering presentation, the students in the experimental class on the immediate posttest of the three lexical words (noun, verb, and adjective), from the mean value, there is a difference in the scores of the three, and the scores of the noun part are slightly higher than the adjective scores, and the difference with the scores of the verb part is more obvious, and the mean score of the noun part is 12.089.

Table 6 shows the results of one-way ANOVA, which mainly tests the differences in the effect of thematic clustering presented in the acquisition of words with different lexical properties. According to the results of the one-way ANOVA, it can be seen that there is a significant difference between the scores of nouns, verbs as well as adjectives ($F = 4.398$, $p = 0.012 < 0.05$). That is, there is a significant difference between the students' performance on different lexical words under the thematic clustering presentation.

Table 7 shows the results of multiple comparisons on the differences in the effects of acquisition of different lexical words with thematic clustering presentation. The scores on the noun part were significantly higher than those on the verb and adjective; the scores on the adjective part were slightly higher than those on the verb, but there was no significant difference between the two (SD = 0.4945, $p = 0.516 > 0.05$), which means that the nouns were best learned under the thematic clustering presentation, while the adjectives and verbs were similarly different. *. The significance level for the difference in means is 0.05.

Table 5. Descriptive statistical analysis of the results of the three words.

	Mean value	Standard deviation	Standard error mean	95%LL	95%UL	Minimum value	Maximum value
Noun score	12.089	2.2806	0.3609	11.348	12.806	8	16
Verb grade	10.689	2.1658	0.3456	9.985	11.369	5	15
Adjective achievement	11.023	2.1856	0.3451	10.3.6	11.699	6	14
Total	11.267	2.2107	0.3505	10.835	11.666	6.33	15.33

Table 6. Analysis of single factor variance.

	Sum of squares	df	Mean square	F	Sig.
Between groups	42.598	2	21.451	4.398	0.012
Within group	571.254	115	4.886		
Total	614.568	116			

Table 7. Multiple comparisons of the results of the three words.

(I)lexical	(J)lexical	Mean	Standard	Sig.	95%CI	95%CI

		difference (I-J)	error mean		Lower	Upper
Noun score	Verb grade	1.4002	0.4945	0.005	0.422	-2.396
	Adjective achievement	1.0756*	0.4945	0.035	0.095	2.058
Verb grade	Noun score	-1.4002	0.4945	0.005	-2.396	-0.422
	Adjective achievement	-0.3254	0.4945	0.516	-1.308	0.659
Adjective achievement	Noun score	-1.0756*	0.4945	0.035	-2.058	-0.095
	Verb grade	0.3254	0.4945	0.516	-0.659	1.308

5.3. The Effect of Thematic Clustering Presentation on Students' Vocabulary Memory Retention

Same as the immediate test, the author conducted four vocabulary delayed posttests and finally took the average score as the students' vocabulary scores on the delayed posttests. SPSS 25.0 was used to conduct independent samples t-test on the delayed post-test vocabulary scores, and the results are shown in Tables 8 and 9.

The experimental class was significantly higher than the control class in vocabulary delayed posttest scores, and the difference between the means of the two groups was 3.944. Table 9 shows the independent samples t-test, and the Levene's variance test showed that the probability of significance (Sig) value was 0.776, which was greater than 0.05, which indicated that the variances of the two classes in the variable of vocabulary delayed posttest scores were equal, and that there was a significant difference in vocabulary scores of the experimental class and the control class in the delayed posttest. That is, there is a significant difference between the vocabulary list presentation and the thematic clustering presentation on the students' vocabulary long-term memory effect, and the thematic clustering presentation is better than the vocabulary list presentation in terms of long-term memory effect.

Table 8. The result of the time delay.

	Class	Number	Mean value	Standard deviation	Standard error mean
Vocabulary achievement	Laboratory class	45	29.356	5.9456	0.9456
	Control class	45	25.412	6.0153	0.9263

Table 9. The independent sample t test was measured after the delay.

	Variance equivalence test		The average value of t is tested						
	F	Sig.	t	df	Sig.(2-tialed)	Mean Difference	Std.Error Difference	95%CI Lower	95%CI Upper
Assumed Equal variances	0.085	0.776	3.056	90	0.004	4.033	1.306	1.405	6.675

Not Assumed Equal variances			3.056	89.4856	0.004	4.033	1.306	1.405	6.675

A comparison of the mean differences between the vocabulary immediate posttest scores and delayed posttest scores for the experimental and control classes is shown in Table 10.

The mean difference between the immediate posttest scores and the delayed posttest scores of the experimental class is 4.43, and the mean difference between the immediate posttest scores and the delayed posttest scores of the control class is 5.80.The scores of both classes show a decreasing trend in the memorization of vocabulary knowledge, but the mean difference between the two posttest scores of the experimental class is smaller than that of the control class, i.e., under the thematic clustering presentation method, the vocabulary knowledge of the students is slower to be forgotten, and vocabulary memorization retention is more effective than that of the control class. Good.

Table 10. The mean difference of the result is instantaneous and delayed.

Class	Posttest	Mean	Mean difference
Laboratory class	Immediate postmeasurement	33.79	4.43
	Postdelay survey	29.36	
Control class	Immediate postmeasurement	30.99	5.80
	Postdelay survey	25.19	

6. Conclusion

This paper designs a business English vocabulary memorization strategy based on K-means clustering analysis algorithm. By organizing and analyzing the experimental data obtained from business English vocabulary clustering and teaching experiments, this study draws the following conclusions:

(1) After K-means clustering Business English words have shown a tendency to hold together, more clusters gathered around the diagonal. The vocabulary representative points before clustering are scattered irregularly in the plane of the coordinate axes. After clustering, the vocabulary points appear clustering characteristics, gathered around the diagonal line, the clustering effect is more satisfactory, indicating that K-means clustering has a better classification effect on the clustering of business English words, and it can extract the public characteristics of different groups of vocabulary.

(2) The business English vocabulary presentation of thematic clustering is more helpful to improve students' business English vocabulary test scores compared with the ordinary presentation. There is a significant difference between the immediate posttest scores of the vocabulary test scores of the experimental class and the control class. The learning effect on nouns and verbs is more obvious, and the effect on nouns is the most significant. Students' forgetting speed is slower under the business English vocabulary presentation of topic clustering, and the vocabulary test scores of the experimental class on the delayed posttest are significantly better than those of the control class, which helps students to maintain their vocabulary memorization effect.

About the Author

Xuefeng Li (born in Oct., 1977), male, Han nationality, native of Lingbi, Anhui, PRC, holds a doctoral degree and works as a lecturer. His main research directions include theoretical linguistics, formal syntax, business English teaching and research, etc.

References

1. Nychkalo, N., Jinba, W., Lukianova, L., Paziura, N. V., & Muranova, N. (2020). Use of task-based approach in teaching vocabulary to business English learners at university. Advanced education, 7(16), 98-103.

2. Wang, Y. H. (2014). Developing and evaluating an adaptive business English self-learning system for EFL vocabulary learning. Mathematical Problems in Engineering, 2014(1), 972184.
3. Zhang, H., Song, W., & Huang, R. (2014). Business English vocabulary learning with mobile phone: A Chinese students' perspective. International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT), 4(2), 46-63.
4. Khamees, K. S. (2016). An evaluative study of memorization as a strategy for learning English. International Journal of English Linguistics, 6(4), 248-259.
5. Pitukwong, K., Soranasathaporn, S., Thanathiti, T., & Engchuan, K. (2014). An investigation of vocabulary learning strategies employed by high and low English proficiency students based on selected vocabulary from a business corpus. Asian International Journal of Social Sciences, 14(1), 106-119.
6. Yan, L., & Yang, Y. (2016). Examining business English majors' business vocabulary knowledge development. The Asian ESP Journal, 12(3), 40-69.
7. Muyassyarah, S. N., Luhriyani, S., & Asfah, I. (2023). Correlation between students'business english vocabulary mastery and their reading comprehension (a study at business english communication students). International Journal of Business English and Communication, 1(3), 85-88.
8. Kovalenko, Y. (2024). Effective techniques for developing advanced vocabulary skills in English language. Teaching Languages at Higher Educational Establishments at the Present Stage. Intersubject Relations, (44), 60-76.
9. Romadhon, R., & Yuliyanti, N. (2025). Exploring the use of chatbots in business english vocabulary learning: students'views and experiences. Judika (jurnal pendidikan unsika), 13(1), 75-98.
10. Run, W., & Weina, L. (2024). Vocabulary thresholds study for business English major based on the range corpus. International Journal of Education and Humanities, 4(1), 1-16.
11. Suchanova, J., Šliogerienė, J., & Mockienė, L. (2017). A paradigm shift in teaching business English vocabulary. Journal of Teaching English for Specific and Academic Purposes, 5(2), 247-257.
12. Abduh, A., & Rosmaladewi, R. (2017). Taking the Lextutor on-line tool to examine students' vocabulary levelin business English students. World Transactions on Engineering and Technology Education, 15(03), 283-286.
13. Malisa, W., Aeka, A., & Lieungnapar, A. (2023, March). influence of contexts on defining business english vocabulary. in international academic multidisciplinary research conference in fukuoka 2023 (pp. 190-197)
14. Al-Qaysi, F. H., & Shabdin, A. A. (2016). Vocabulary memorization strategies among Arab postgraduate English foreign language learners. Advances in language and Literary Studies, 7(5), 184-196.
15. Huang, Z. (2022, November). Exploring effective teaching strategies for business English beginners. In 2022 International Conference on Science Education and Art Appreciation (SEAA 2022) (pp. 635-642). Atlantis Press.
16. Shavkidinova, D. (2022). Teaching English vocabulary through vocabulary classification techniques. European International Journal of Multidisciplinary Research and Management Studies, 2(10), 189-201.
17. Al-Faris, S., & Jasim, B. Y. (2021). Memory strategies and vocabulary learning strategies: Implications on teaching and learning vocabulary. Journal of Humanities and Social Sciences Studies, 3(10), 11-21.
18. Komiljonovna, R. S. (2024). Teaching English Vocabulary for Specific Purposes: A Comprehensive Guide. Research Focus, 3(9), 105-109.
19. Wierzchoń, S. T., & Kłopotek, M. A. (2018). Modern algorithms of cluster analysis. Springer.
20. Saraçli, S., Doğan, N., & Doğan, İ. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. Journal of inequalities and Applications, 2013(1), 203.
21. Backhaus, K., Erichson, B., Gensler, S., Weiber, R., & Weiber, T. (2025). Cluster analysis. In Multivariate analysis: An application-oriented introduction (pp. 461-538). Wiesbaden: Springer Fachmedien Wiesbaden.
22. Aggarwal, C. C. (2018). An introduction to cluster analysis. In Data clustering (pp. 1-28). chapman and hall/crc.
23. Jaeger, A., & Banks, D. (2023). Cluster analysis: A modern statistical review. Wiley Interdisciplinary Reviews: Computational Statistics, 15(3), e1597.
24. Rahimi, H. (2014). The effect of method of vocabulary presentation (code-mixing, thematic clustering, and contextualization) on L2 vocabulary recognition and production. Procedia-Social and Behavioral Sciences, 98, 1475-1484.
25. Khoii, R., & Sharififar, S. (2013). Memorization versus semantic mapping in L2 vocabulary acquisition. ELT journal, 67(2), 199-209.
26. Jang, H. J. (2014). The Effects of Semantic Clustering on EFL Young Learners' Vocabulary Learning. English Teaching, 69(3).
27. Wu, Q. (2014). A rote strategy in memorizing vocabulary for ESL learners. Procedia-Social and Behavioral Sciences, 143, 294-301.
28. Wang, X., Liu, Q., Pang, H., Tan, S. C., Lei, J., Wallace, M. P., & Li, L. (2023). What matters in AI-supported learning: A study of human-AI interactions in language learning using cluster analysis and epistemic network analysis. Computers & Education, 194, 104703.
29. Shahzad, S. K., Sarwat, S., & Kanwal, S. (2023). Use Of Semantic and Clustering Methods: for Teaching English Vocabulary At Elementary Level. International" Journal of Academic Research for Humanities", 3(4), 74-84.
30. Ambarwati, N., & Saragih, G. (2021). Clustering technique and vocabulary mastery towards students'writing skill in recount text. Inference: Journal of English Language Teaching, 4(1), 10-15.

31. Bakri, A. F. (2022). STUDENTS'SELF-STRATEGIES IN MEMORIZING ENGLISH VOCABULARY. NUSRA: Jurnal Penelitian dan Ilmu Pendidikan, 3(2), 191-199
32. Mohamad, N. Z., Hashim, Z., Parjan, H. W., Shukor, S. N. E. A., Rajagopal, K., & Hashim, H. (2021). Students' perception of using memory strategies training for vocabulary development. International Journal of Academic Research in Business and Social Sciences, 11(7), 315-328.
33. Jingzhi, Z., & Lin, W. (2019). Analysis of the Vocabulary Characteristics and Teaching Strategy of Business English Based on Corpus. Journal of Social Sciences Studies, 3, 286-289.
34. Van, N. V. P., & Linh, D. Q. (2024). Vocabulary Learning Strategies Used By The First-Year Students Of The Conducted-In-English Program At The University Of Economics And Business Administration. International Journal Of All Research Writings, 5(11), 36-38.
35. Keemthong, S. (2022, September). Shortcuts to Memorizing English Academic Vocabulary A Case Study of English-Major Students in the Self-Regulated Learning Environment. In Proceedings 5th International Conference of Sustainable Development (ICSD) 2021 (pp. 93-100).
36. Jaikrishnan, S., & Ismail, H. H. (2021). A review on vocabulary learning strategies used in learning English as a second language. International Journal of Academic Research in Business and Social Sciences, 11(9), 297-309.
37. Aminatun, D., & Oktaviani, L. (2019). Using "Memrise" To Boost English For Business Vocabulary Mastery: Students'Viewpoint. Proceedings Universitas Pamulang, 1(1).
38. Martins, H. F., & Ferro, M. J. (2021). Vocabulary learning strategies: The case of English for business and financial reporting. International Journal of Language and Literary Studies, 3(3), 316-330.
39. Le, H. S., & Trinh, M. L. (2024). An investigation of vocabulary learning strategies of ESP students. International Journal of TESOL & Education, 4(1), 1-17.
40. Kohnke, L., Zhang, R., & Zou, D. (2019). Using mobile vocabulary learning apps as aids to knowledge retention: Business vocabulary acquisition. Journal of Asia TEFL, 16(2), 683.
41. Heng, K. (2020). 2. Effective English vocabulary learning strategies: A research summary. Edited by Kimkong Heng Sopheap Kaing Vutha Ros Koemhong Sol, 12.
42. Andari, I. A. M. Y., Wiguna, I. B. A. A., & Arini, N. M. (2022). The use of flashcards teaching strategy in recalling English vocabulary. Yavana Bhasha: Journal of English Language Education, 5(1), 4-13.
43. Chatzi Anna V. (2025). Understanding the independent samples t test in nursing research. British Journal of Nursing,34(1),56-62.https://doi.org/10.12968/BJON.2024.0133.
44. Fei Chen, Zhenling Zhang, Yangli Jia, Ruchao Jia, Haitao Wang & Xinyu Cao. (2025). Research on the similarity calculation of short text in the terminology domain based on siamese BERT model. Scientific Reports,15(1),36954-36954.https://doi.org/10.1038/S41598-025-20908-8.
45. Dias Leonardo A., Ferreira Joao C. & Fernandes Marcelo A. C.. (2020). Parallel Implementation of K-Means Algorithm on FPGA.IEEE Access,8,41071-41084. https://doi.org/10.1109/access.2020.2976900.