CONSTRUCTING A TEXT-MINING BASED ENGLISH VOCABULARY LEARNING LIST-A CASE STUDY OF COLLEGE ENTRANCE EXAMINATION IN TAIWAN

*Yi-Ning Tu, Yu-Fang Lin and Jou-Cuei Chan

Department of Statistics and Information Science, College of Management, Fu Jen Catholic University, New Taipei City 24205, Taiwan (R.O.C.)
eniddu@gmail.com

ABSTRACT

This study applied text mining techniques, machine learning approaches and statistical methods to construct a predictive model of a prioritized English vocabulary list to help nonnative English speakers prepare for college entrance English exams. Developing a method for efficiently learning English vocabulary in a limited time is an import issue. This study suggests that highly relevant and frequently repeated test items should be learned first. Although the College Entrance Examination Center (CEEC) in Taiwan has provided an approximately 7,000-word vocabulary list, the list's suitability requires verification. Furthermore, this study constructed a vocabulary learning process model to establish a prioritized English vocabulary list for future examinees. Experimental results show that the proposed model can achieve a 78% hit ratio, which is higher than the 69% of the CEEC's provided list.

KEYWORDS

Text mining, Machine learning, Vocabulary Learning Process Model, College entrance exam, EFL (ESL).

1. INTRODUCTION

1.1 Research background

Learning English is critical, especially in countries where English is the first foreign language (EFL, English as Foreign Language) or (ESL, English as the Second Language). In Taiwan, the English score on the College Entrance Examination Center (CEEC) examination is one of the key indices for college admission. However, Taiwan's current college entrance examinations conform with the "one curriculum guidelines, multiple versions of textbooks" policy. Thus, most students consider the English test too difficult to prepare for, because the use of multiple English textbooks may necessitate studying a more diverse range of articles and vocabulary words than that required in other disciplines. Furthermore, Jia et al. (2012) stated that students cannot retain memorized English vocabulary for a long time. Consequently, helping students and examinees study English vocabulary in a strategic manner is imperative.

English literacy is a crucial index used to gauge English ability. It is determined by word ability (Hinkel, 2006; Schmitt, 2000). Word ability is defined as the volume of vocabulary that a learner understands and can apply. Chen and Chung (2008) claimed that because sentences are composed of words, expanding the vocabulary improves a learner's English fluency. Furthermore, with a large vocabulary, a student easily understands the meanings of sentences in an article. Astika (1993) and other researchers such as Laufer and Nation (1995) have indicated that having a

DOI: 10.5121/ijdkp.2016.6604

greater word ability improves learners' reading and writing skills. Lin (2007) reported that word ability and writing skill are positively correlated. Therefore, if students want to perform well in examinations, the highest priority is to increase their word ability.

Vocabulary is the basis of a language. When preparing for a test with an extremely wide scope, prioritizing vocabulary words is an efficient approach to achieving higher scores. If examinees have an ordered vocabulary list to learn, they can review or practice words at their personal pace and degree. In other words, examinees can learn many words in a limited range at once, but not an entire language's words in alphabetical order.

1.2. Research issues

(1) The CEEC provided an approximately 7,000-word vocabulary list to Taiwanese senior high schools in 2000 (Zheng, 2002) to help examinees. However, the present suitability of the vocabulary list, which was compiled many years ago, is doubtful. The current study not only explored the relationship between the past exam items in each year and the vocabulary list provided by the CEEC but also examined which years' exams were the most consistent with the provided list.

(2) It was theorized that essential concepts will be tested repeatedly in future examinations and that knowing the items of past examinations can help examinees prepare for the vocabulary anticipated in future tests. Therefore, this study investigated the correlations and regularities between the provided vocabulary list and the past items in each year's examination. The results can provide examinees with a reference list for studying for college entrance examinations.

(3) Learners must first focus on simple words before learning the more difficult ones. Hence, this study categorized words according to the stage of learning English that they belong to. For example, the categories include the basic 1,000 words learned in elementary and junior high school (E&J), words from high school textbooks, and words used in test items on past examinations. This study used conditional probability to construct a vocabulary list for helping examinees increase their examination scores.

(4) Finally, the proposed vocabulary list was prioritized by the words' relevancy and probability of appearing on another examination. Therefore, within the limited time for preparation, examinees can decide to review the words that have a higher probability than that of others to appear on future examinations.

2. RELATED WORKS

Learning English has recently attracted considerable research interest. Some papers have discussed language units in learning English. Other related papers have discussed teaching English strategically to prepare for English examinations. The structure of Section 2 is as follows. In Section 2.1, this paper reviews English examinations and teaching strategies in Taiwan. Section 2.2 describes the research unit used for vocabulary. Section 2.3 shows how the related work applied computer technologies for helping EFL (English as Foreign Language) students. Section 2.4 illustrates the approaches to assessing a proposed model's performance.

2.1. English examinations and teaching

In recent years, numerous researchers have discussed English teaching strategies and the CEEC in Taiwan. Many papers have described the obvious trend of the testing of English determining the teaching of it. The CEEC examination has been divided into two stages since 2002. The first stage is the Scholastic Achievement Test (SAT), which is held in the end of February, and the second stage is the Department Required Test (DRT), which is held in the beginning of July. The SAT, the DRT, and an interview exam are the three base scores used to apply to colleges.

The current teaching and examination policies of the CEEC are based on the practice of using many textbooks for one guideline. In other words, the course outline is identical, but the textbooks vary. This policy may require the examinees to study a considerable amount of information, especially in the subject of English. The vocabularies of various textbooks are too diverse for examinees to memorize. To help examinees overcome this difficulty, the CEEC entrusted a research team to propose a 7,000-word vocabulary list for senior high school students (Zheng, 2002). However, because this list was published more than 10 years ago, the suitability of its terms has become uncertain.

Widely used vocabulary lists were considered in this study. The Ministry of Education in Taiwan announced the E&J. The General English Proficiency Test also provides a list for examinees to reference. However, these vocabulary lists are not specifically purposed for the CEEC examination, and the excessive information they contain may lead to information overload among examinees.

In a discussion of CEEC English examinations, Fan (2008) concluded that the Department Required English Test (DRET) is more difficult than the Scholastic Achievement English Test (SAET) regarding vocabulary and similar to it regarding format. Chou (2009) found that from 2006 to 2008, the proportion of difficult items increased in the DRET, and although the examinees performed well on inference items, such as relative words and conjunction elements, they still had problems with items of fixed usage such as preposition, phrases, and reiteration. In summary, the DRET is generally more difficult than the SAET in various aspects.

Regarding English teaching, applying digital or information technology has become a new trend. Jia, Chen, Ding, and Ruan (2012) used a Moodle model to improve the examination scores of students. Moodle is a widely used, free, and open-source course management system. The results indicated that after finishing all the courses, the students in an experimental group learned vocabulary more effectively than the students in a control group did. AbuSeileek (2011) observed substantial differences in text memory and vocabulary acquisition after teaching with hypermedia annotations between learners who used multimedia and those who used hyperlinks. Furthermore, Chen and Chung (2008) proposed a system that considered a user’s interests, preferences, and abilities to provide an appropriate degree of course materials and thus increase learning. In summary, digital or information technology can help students learn English efficiently.

Previous research has focused on English learning interests and discussed the effectiveness of tools such as information systems in helping examinees and students. However, the current study proposes a perspective that considers the study of prioritized vocabulary. Because a vocabulary list is crucial, especially for examinees who must prepare for the CEEC examination efficiently, this study established a vocabulary learning process model (VLPM) for the CEEC examination to determine the priority and frequency of examination of words from each learning stage.

2.2. Research unit: vocabulary

Many related studies define the research unit before starting a research experiment. In this study, the smallest unit in English learning is a word. However, there are two approaches to counting words: by word family and by lemma (Yang, 2006). In the first approach, the word family unit contains the base word, winding words, and derivation words. For example, read, reads, reading, readable, and readability are considered to belong to the same unit, with read being the base word. Reads and reading are classified as winding words. Readable and readability are classified as derivation words. The classification logic is shown in Fig. 1.

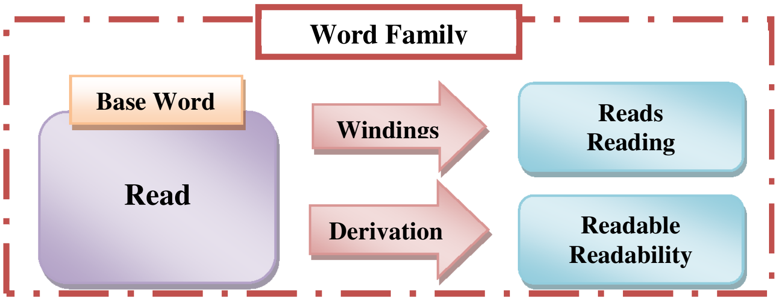


Fig. 1. The unit of word family

By contrast, the lemma unit contains the base word. windings word, and lemma unit equivalents to the word. However, each derivative word is regarded as another unit. For example, read, reads, and reading are classified as the same unit, but read, readable, and readability are classified as three separate units, as shown in Fig. 2. The current study used the lemma as the vocabulary research unit because the textbooks in Taiwan use the lemma as the word unit.

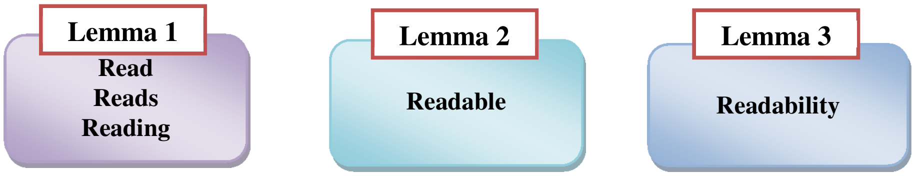


Fig. 2. The unit of lemma

2.3. Computer technologies applied for EFL students

There are a lot of related works which applied the computer technologies for helping EFL students. For example, Hsu (2008) suggested an online personalized English learning recommender system capable of providing ESL students. Hsu, Hwang, & Chang (2013) developed a personalized recommendation-based mobile language learning approach for guiding EFL students and had good performance. Huang and his colleagues (2012) develop a ubiquitous English vocabulary learning (UEVL) system to assist students in experiencing a systematic vocabulary learning process in which ubiquitous technology is used to develop the system, and video clips are used as the material. Smith and his colleagues (2014) investigated how Chinese undergraduate college EFL students to learn new vocabulary with inference-based computer games and also had good performance. Sandberg, Maris and Hoogendoorn (2014) suggested that added value of a gaming context and intelligent adaptation for a mobile learning can improve the students' performance in vocabulary acquisition and in ordinary test. Wu (2015) developed an app and studied its effectiveness as a tool in helping EFL college students learn English vocabulary. That study proved that using the program has a higher performance than control group in acquiring new vocabulary. To our best knowledge, it seems no related works that discuss the teaching for learning strategies for EFL students who have to prepare the exam in Taiwan.

2.4. Approaches to performance assessment

Any evaluation mechanism must have a fair and reasonable assessment procedure (Lin, 2008). The common assessment indices in prediction models are precision, recall, and the F-measure. Van Rijsbergen proposed the F-measure in 1979 according to the harmonic mean of precision and recall. Precision considers all documents and only the target results returned by the system. Recall is the fraction of the documents relevant to the query that are successfully retrieved. Previous studies using the ontology approach (Boonchom & Soonthornphisaj, 2012; Zheng, Chen, &

Jiang, 2012), semantic approach (François & Christine, 2009; Li, Yang, & Park, 2012; Osman, Salim, & Binwahlan, 2012), or evolving methods (Luo, Li, & Chung, 2009; Uguz, 2011) have used these three indices to assess the effectiveness of a prediction model.

This current research used the three indices to measure the accuracy of the proposed VLPM. Accordingly, a confusion matrix was defined and is shown in Table 1:

Table 1. Confusion matrix

		Predicted	
		Yes	No
Actual	Yes	*a*	*b*
	No	*c*	*d*

Where

- *a* is the number of correct predictions that a vocabulary word appears in the examination;
- *b* is the number of incorrect predictions that a vocabulary word does not appear in the examination;
- *c* is the number of incorrect predictions that a vocabulary word appears in the examination; and
- *d* is the number of correct predictions that a vocabulary does not appear in the examination.

$$Precision\ (P) = \frac{\text{number of correct predictions that a vocabulary word appears in the examination}}{\text{number of predictions of the appearance of a vocabulary word in the examination}} = \frac{a}{a+c} \quad (1)$$

$$Recall\ (\mathrm{R}) = \frac{\text{number of correct predictions that a vocabular word appears in the examination}}{\text{number of appearances of a vocabulary word in the examination}} = \frac{a}{a+b} \quad (2)$$

The F-measure is a metric of a model's performance (Lee et al., 2007). The closer the F-measure value is to 1, the more accurate the predicted model is. This formula is shown in Equation (3).

$$F - measure = \frac{2 \times P \times R}{(P+R)} \quad (3)$$

- P: Precision
- R: Recall

Cosine similarity assesses the similarity of two documents. The closer the value is to 1, the higher the ratio of similar content is between the two documents. Cosine similarity was used in the current study to indicate the similarity of two examinations given two vectors $\overrightarrow{D_1}$ and $\overrightarrow{D_2}$. The formula is provided in Equation (4):

$$cos\left(\overrightarrow{D_1}, \overrightarrow{D_2}\right) = \frac{\overrightarrow{D_1} \times \overrightarrow{D_2}}{\left|\overrightarrow{D_1}\right| \times \left|\overrightarrow{D_2}\right|},\ \left(0 \le cos\left(\overrightarrow{D_1}, \overrightarrow{D_2}\right) \le 1\right) \quad (4)$$

Where

- $\overrightarrow{D_1}$ indicates the document D_1 vector; and
- $\overrightarrow{D_2}$ indicates the document D_2 vector.

3. Concepts of the VLPM

To help examinees prepare for the CEEC English examination, vocabulary lists of previous exam items were collected and used as the study data sets. Categories were then used to classify these words by difficulty level, and the conditional probability and joint probability mass functions were applied as the computing method to propose a VLPM of the CEEC English examination. Section 3 is organized as follows. Section 3.1 describes the concepts of the VLPM design.

Section 3.2 explains a suitability check module used to validate the suitability of the vocabulary list suggested by the CEEC. Section 3.3 describes a pattern extraction module that was used to examine whether there was any relationship between past examinations and to extract the most frequently tested examination words. Finally, Section 3.4 details the VLPM for prioritizing all of the categorized words for the examinees.

3.1. Concepts of the VLPM design

This study can be divided into four major parts that illustrate and explain the development process of the proposed VLPM. It was assumed that the content of past examinations varies in each year, but the key concepts are tested repeatedly. The four major parts are as follows.

- Comparing the suitability of past exam items with that of the CEEC's suggested vocabulary list for preparing for future examinations.
- Determining the regularity or correlation between past exam items to predict future exam vocabulary.
- Using different vocabulary lists to investigate the properties of the predicted vocabulary list and establish the VLPM.
- Prioritizing the vocabulary to construct an English vocabulary list for preparation for future CEEC examinations. Fig. 3 shows the four major modules of this study.

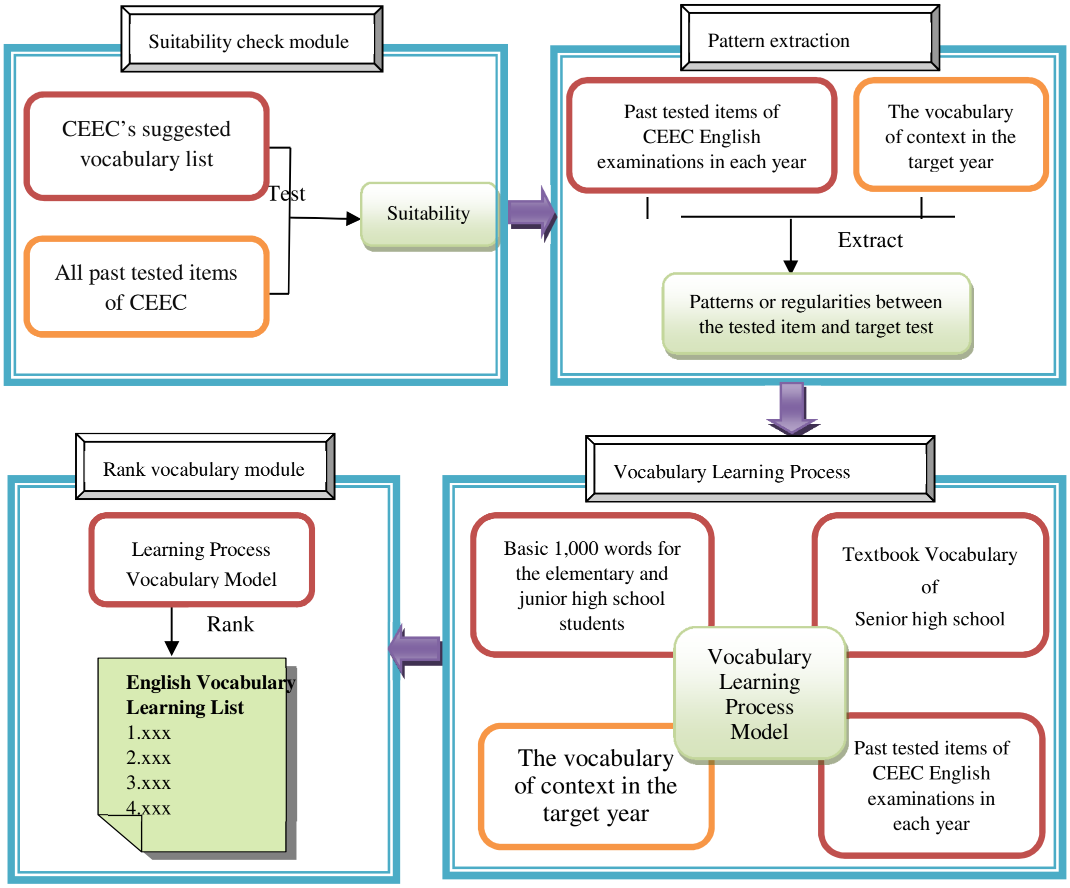

Suitability check module
CEEC's suggested vocabulary list
Test
Suitability
All past tested items of CEEC
Pattern extraction
Past tested items of CEEC English examinations in each year
The vocabulary of context in the target year
Extract
Patterns or regularities between the tested item and target test
Rank vocabulary module
Learning Process Vocabulary Model
Rank
English Vocabulary Learning List
1.xxx
2.xxx
3.xxx
4.xxx
Vocabulary Learning Process
Basic 1,000 words for the elementary and junior high school students
Textbook Vocabulary of Senior high school
Vocabulary Learning Process Model
The vocabulary of context in the target year
Past tested items of CEEC English examinations in each year

Fig. 3. The conceptual of the Vocabulary Learning Process Model

3.2. Suitability check module

The suitability check module compares the suitability of the CEEC's suggested vocabulary list with that of all of the past exam items. Because it is logical that essential concepts are retested repeatedly, examinees can focus on the past exam items to prepare for the examination in their target year. Consequently, all the items of past examinations were collected for this study as a vocabulary list to validate the theory that relevant concepts are retested repeatedly. The items on the target examination of the nth year were used as the testing data, and the items on past examinations from the 1th to (n -1)th years were used as the training data. The goal of this module was to determine whether the past exam items are a more suitable reference for preparing for the CEEC English examination than the vocabulary list proposed by the CEEC.

This study used the lemma as the measurement unit to calculate the numbers of vocabulary words on past examinations and words on the CEEC's vocabulary list. The F-measure value was used as the performance index. The higher F-measure value among the two word lists was used to determine the most suitable list, as shown in Fig. 4.

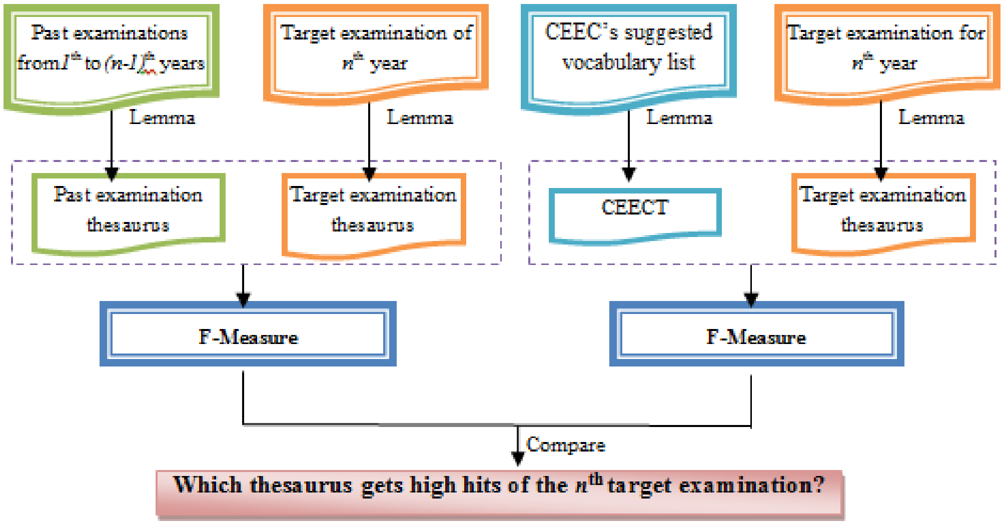

Past examinations from 1st to (n-1)th years
Target examination of nth year
CEEC's suggested vocabulary list
Target examination for nth year
Lemma
Lemma
Lemma
Lemma
Past examination thesaurus
Target examination thesaurus
CEECT
Target examination thesaurus
F-Measure
F-Measure
Compare
Which thesaurus gets high hits of the nth target examination?

Fig. 4. The experimental design of suitability check module

3.3. Pattern extraction module

The goal of the pattern extraction module was to determine the correlations and the regularities between the vocabulary of past examinations and the content in the target year. According to the theory that essential concepts are retested repeatedly, it was assumed that there were patterns embedded in past examination content. It was further theorized that exam committees may reference recent exam content and thus retest emerging relevant concepts during a period of time. For example, the content in the most recent three years may have some relationships or regularities.

The words of past examinations were quantified separately using lemmas. The list from the first year of the CEEC examination was referred to as the first vocabulary list, and the list from the nth year was referred to as the vocabulary list for the target examination in the nth year. Likewise, the (n-1)th vocabulary list was from the year directly prior to the target year, and so on. The design of the module experiment is shown in Fig. 5.

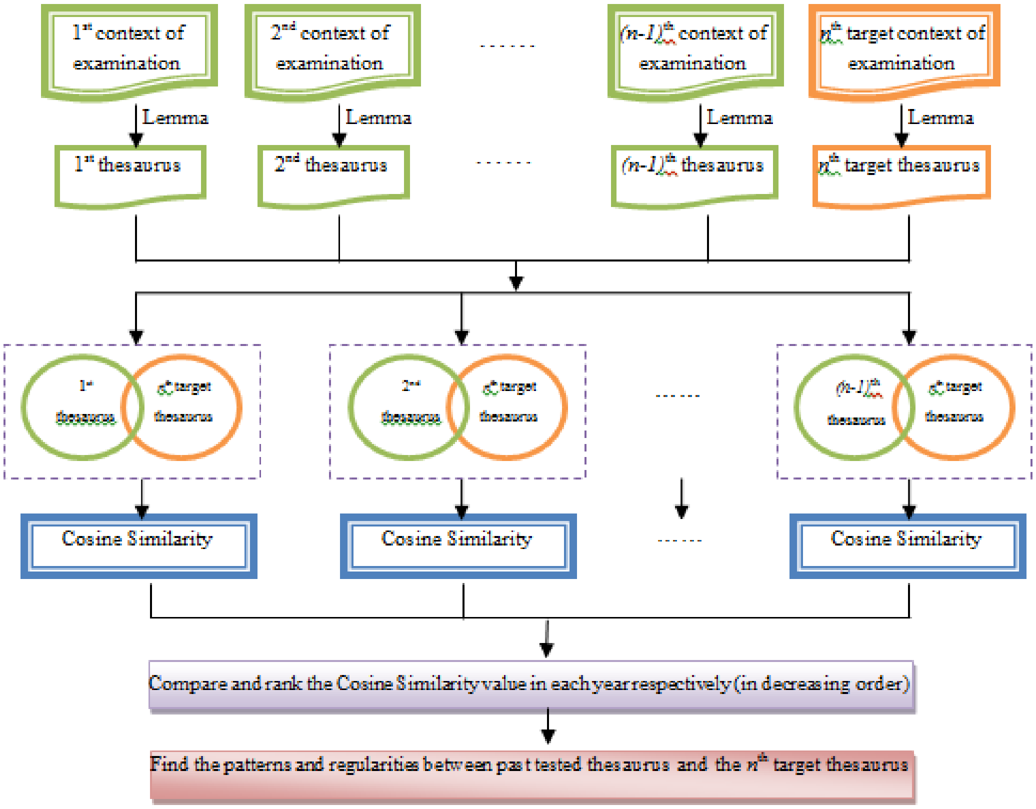

1st context of examination
2nd context of examination
......
(n-1)th context of examination
nth target context of examination
Lemma
Lemma
Lemma
Lemma
1st thesaurus
2nd thesaurus
......
(n-1)th thesaurus
nth target thesaurus
1st thesaurus
nth target thesaurus
2nd thesaurus
nth target thesaurus
......
(n-1)th thesaurus
nth target thesaurus
Cosine Similarity
Cosine Similarity
......
Cosine Similarity
Compare and rank the Cosine Similarity value in each year respectively (in decreasing order)
Find the patterns and regularities between past tested thesaurus and the nth target thesaurus

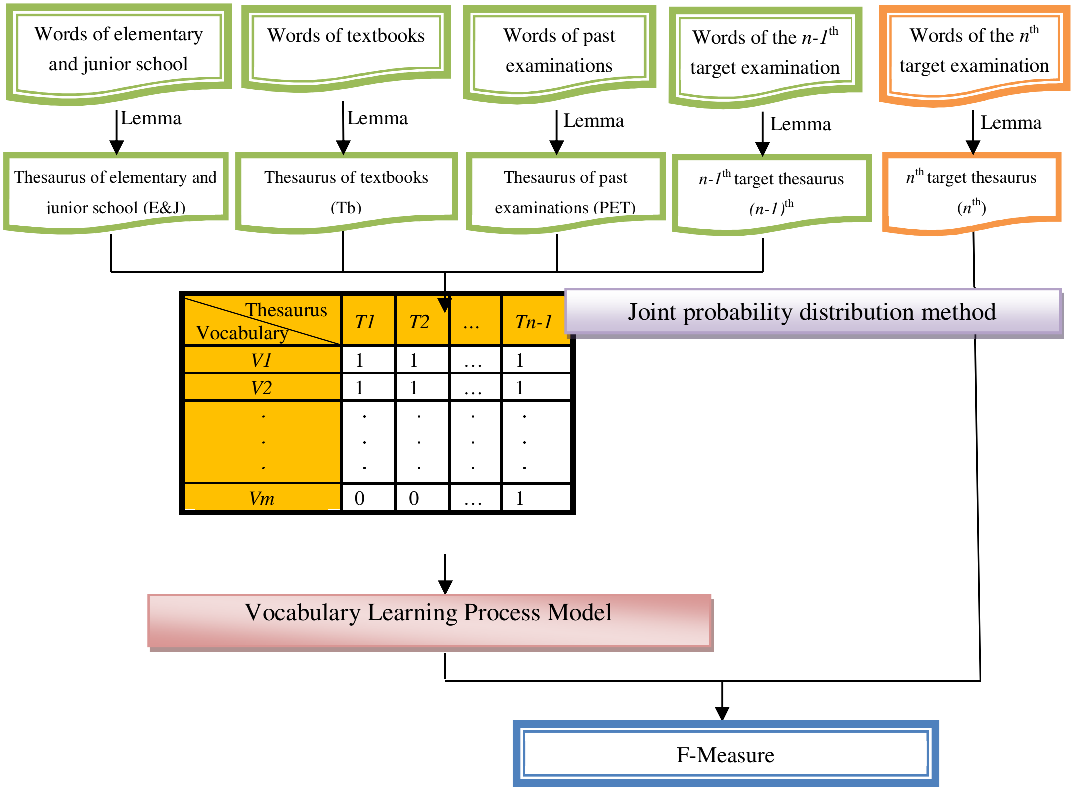
Fig. 5. The experimental design of patterns extraction module

3.4. Vocabulary Learning Process Model

After exploring the patterns between past exam items and the target examination, the vocabulary lists were used for categorizing the large volume of words and prioritizing them within a more detailed range than that provided by the CEEC. The VLPM was used to build 2n modules by using n word lists, and to determine which module had the higher hit ratio for the nth target year of the examination. The conditional probability and joint probability mass functions were used to establish the VLPM. It was assumed that when an examinee starts to prepare for a test, there should be a learning progression from frequently tested words to less frequently tested words. Furthermore, the easy words should be learned earlier in preparation than the more difficult words. Using the learning process of Taiwan English education as an example, this study used different vocabulary lists to categorize words into individual learning stages.

The E&J is learned in elementary school and junior high school. The vocabulary learned from senior high school textbooks was termed the textbook word list (TbT). Furthermore, because past examination items may appear on the future target examination, all items on past examinations from the first to (n-2)th years were included as the past test items of the CEEC word list (PCEECT). The (n-1)th year was defined as the target examination and denoted as the target word list (Target T). The nth year examination content was used as the validating data to examine the accuracy of the proposed model.

The E&J, TbT, and PCEECT were collected to construct the (n-1)th VLPM. The (n-1)th vocabulary list proposed by the (n-1)th VLPM was then used to test the nth test content and validate the model accuracy. The process of establishing the VLPM is illustrated in Fig. 6.


Words of elementary and junior school
Words of textbooks
Words of past examinations
Words of the n-1th target examination
Words of the nth target examination
Lemma
Lemma
Lemma
Lemma
Lemma
Thesaurus of elementary and junior school (E&J)
Thesaurus of textbooks (Tb)
Thesaurus of past examinations (PET)
n-1th target thesaurus (n-1)th
nth target thesaurus (nth)
Joint probability distribution method
Vocabulary Learning Process Model
F-Measure

Thesaurus / Vocabulary	*T1*	*T2*	…	*Tn-1*
V1	1	1	…	1
V2	1	1	…	1
. . .	. . .	. . .	. . .	. . .
Vm	0	0	…	1

Fig. 6. The establishing process of the vocabulary learning process model (VLPM)

4. Establishing The VLPM

This section describes how the VLPM was established. The steps were (1) establishing a two-dimensional Boolean matrix; (2) understanding the distribution of every module; (3) calculating the probability of each node; (4) computing the conditional probability of each node; (5) computing the joint probability of each module; (6) calculating the ratio that appears in the Tn vocabulary list; (7) prioritizing the vocabulary words to be used in English examinations.

(1) Establishing the two-dimensional Boolean matrix

A vocabulary list (T) and word (V) Boolean matrix was fabricated, as shown in Table 2. Researchers can use the table to determine which words are included in multiple vocabulary lists and therefore more likely to be tested than others.

Table 2. The Boolean matrix by the vocabulary list (*T*) and word (*V*)

vocabulary list (*T*) word(*V*)	T_1	T_2	…	T_i	…	T_{n-1}
V_1	1	0	…	1	…	1
V_2	1	1	…	0	…	0
.	.	.	.	.	.	.
V_a	.	.	.	.	.	.
.	.	.	.	.	.	.
V_{m-1}	0	1	…	1	…	0
V_m	1	0	…	0	…	1

Here,

- T_i is the i^{th} vocabulary list ($I = 1,2,\ldots,n - 1$);
- V_a is the a^{th} word ($a = 1,2,\ldots,m$); and
- a value of 1 indicates that the word appears in the vocabulary list, whereas 0 indicates that the word does not appear in the vocabulary list.

(2) Understanding the probability distribution of every module

After establishing the vocabulary list (*T*) and word (*V*) Boolean matrix, the combination of every vocabulary list, such as T_1-T_2-…-$T_{n-1,}$ was defined as an attributed module (M). Each attributed module was mutually exclusive, and Table 3 shows the marginal probability distribution of each one. The ratio was then computed. If the first word (V_1) in the database appears in the first vocabulary list (T_1), the cell value is 1, as shown in Table 3. The probability of each module is based on Equation (5):

$$P\left(\frac{T_1 - T_2 - \cdots - T_{n-1}}{M_j}\right) = \frac{X_k}{A} \qquad (5)$$

Where,

- n is the target year of the CEEC examination;
- T_i is the i^{th} vocabulary list ($I = 1,2,\ldots,n - 1$);
- T_1-T_2-…-T_{n-1} is the vocabulary list combination;
- M_j is the j^{th} module ($j = 1,2,\ldots,2^{n-1}$);
- X_k is the frequency of M_j ($k = 1,2,\ldots, 2^{n-1}$); and
- A is the total number of words.

Table 3. Marginal probability distribution of each module

Thesaurus Module	T_1	T_2	…	T_i	…	T_{n-1}	Probability
M_1	0	0	…	0	…	0	X_1/A
M_2	0	0	…	0	…	1	X_2/A
.	.	.	…	…	…	.	.
M_j	.	.	…	…	…	.	X_k/A
.	.	.	…	…	…	.	.
$M_{2^{n-1}}$	1	1	…	1	…	1	X_r/A

(3)Calculating the probability of each node

Before calculating the conditional probability of each node, the joint probability values, such as *P(W), P(W, X), P(W, X, Y), and P(W, X, Y, Z),* were calculated for each node as shown in Table 4.

Table 4.Joint probability of $P(T_2, T_1)$

Thesaurus Module	T_1	T_2	$P(T_2, T_1)$
M_1	0	0	x_1/A
M_2	0	1	x_2/A
M_3	1	0	x_3/A
M_4	1	1	x_4/A

(4)Computing the conditional probability of each node

The marginal probability and joint probability of each node were used to compute the conditional probability. This is achieved by using Equation (6), and the results are shown in Table 5.

$$P(X|W) = \frac{P(X,\ W)}{P(W)} \tag{6}$$

Here,

- W is the number of words in T_1; and
- X is the number of words in $T_{2.}$

Table 5.Conditional probability of $P(T_2| T_1)$

$P(T_2)$ $P(T_1)$	$P(T_2/ T_1)$	
	T_2=0	T_2=1
T_1=0	$\frac{\frac{x_{00}}{A}}{P(T_1)}$	$\frac{\frac{x_{01}}{A}}{P(T_1)}$
T_1=1	$\frac{\frac{x_{10}}{A}}{P(T_1)}$	$\frac{\frac{x_{11}}{A}}{P(T_1)}$

(5)Computing the joint probability of each module

The joint probability of each module was computed according to the conditional probability by using Equation 7. The results are shown in Table 6.

$$P(W, X, Y, Z) = P(W) \times P(X|W) \times P(Y|WX) \times P(Z|WXY) \tag{7}$$

Here

- $Y_{j,}$ and N_j are the values of the joint probability in each module ($j = 1,2,\ldots,2^{n-1}$);
- Y_1 is the value of the joint probability distribution that appeared in the T_{n-1}^{th} vocabulary list when the T_1- T_2-...T_{n-2} vocabulary list is 0; and
- N_1 is the value of the joint probability distribution that does not appear in the T_{n-1}^{th} vocabulary list when the T_1- T_2-...T_{n-2} vocabulary list is 0.

Table 6. The joint probability of T_1-T_2- …-T_{n-1}thesaurus

Thesaurus Module	T_1	T_2	…	T_{n-1}	Joint probability of Module
M_1	0	0	…	1	Y_1
	0	0	…	0	N_1
M_2	1	1	…	1	Y_2
	1	1	…	0	N_2
	.	.	…	.	.
.	.	.	…	.	.
.	.	.	…	.	.
.	.	.	…	.	.
	.	.	…	.	.
	.	.	…	.	.
$M_{2^{n-1}}$	0	1	…	1	Y_e
	0	1	…	0	N_f

(6)Calculating the ratio that appears in the T_n vocabulary list

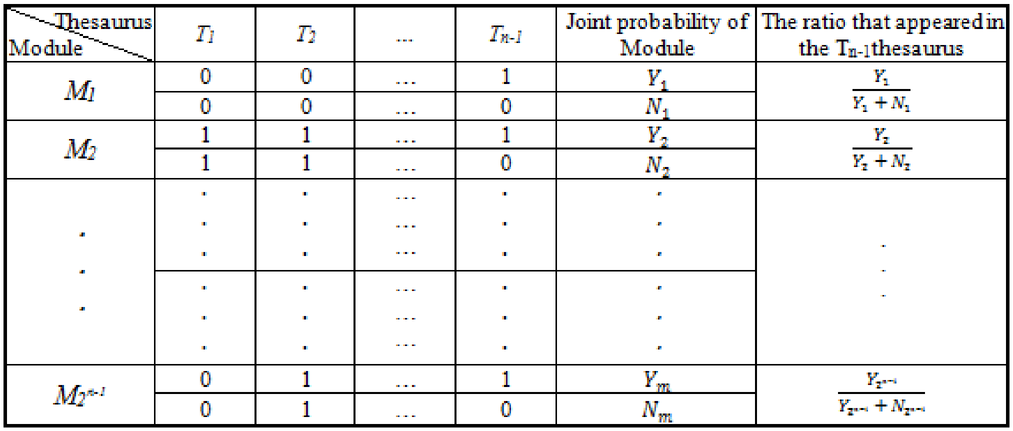
The model T_1- T_2-…T_{n-2} was used to explore the ratio of tested to not tested words for any T_1-T_2-…T_{n-2} vocabulary list combination in the T_{n-1} vocabulary list by using Equation (8). The ratio was ranked from high to low.

$$T_{n-1}\text{word list in which the word appears} = \frac{Y_j}{Y_j + N_j} \qquad (8)$$

Where,

- Y_j is the probability that a word appears in the M_j module ($j = 1,2,\ldots,2^{n-1}$); and
- N_j is the probability that a word does not appear in the M_j module ($j = 1, 2,\ldots,2^{n-1}$).

Table 7.The value of probability that appears in the M_j module



Thesaurus Module	T_1	T_2	…	T_{n-1}	Joint probability of Module	The ratio that appeared in the T_{n-1}thesaurus
M_1	0	0	…	1	Y_1	$\frac{Y_1}{Y_1+N_1}$
	0	0	…	0	N_1	
M_2	1	1	…	1	Y_2	$\frac{Y_2}{Y_2+N_2}$
	1	1	…	0	N_2	
	.	.	…	.	.	
.	.	.	…	.	.	
	.	.	…	.	.	.
.	.	.	…	.	.	.
	.	.	…	.	.	.
.	.	.	…	.	.	
$M_{2^{n-1}}$	0	1	…	1	Y_m	$\frac{Y_{2^{n-1}}}{Y_{2^{n-1}}+N_{2^{n-1}}}$
	0	1	…	0	N_m	

(7)Prioritizing words for preparation for English examinations

The goal of this study was to help examinees memorize a large vocabulary list with a higher hit ratio in a limited time. Hence, the modules were divided using the VLPM into categorized vocabulary lists and the required words were prioritized by the hit ratio and compared with the CEEC's provided list. This provided a suitable vocabulary ranking and English vocabulary learning list.

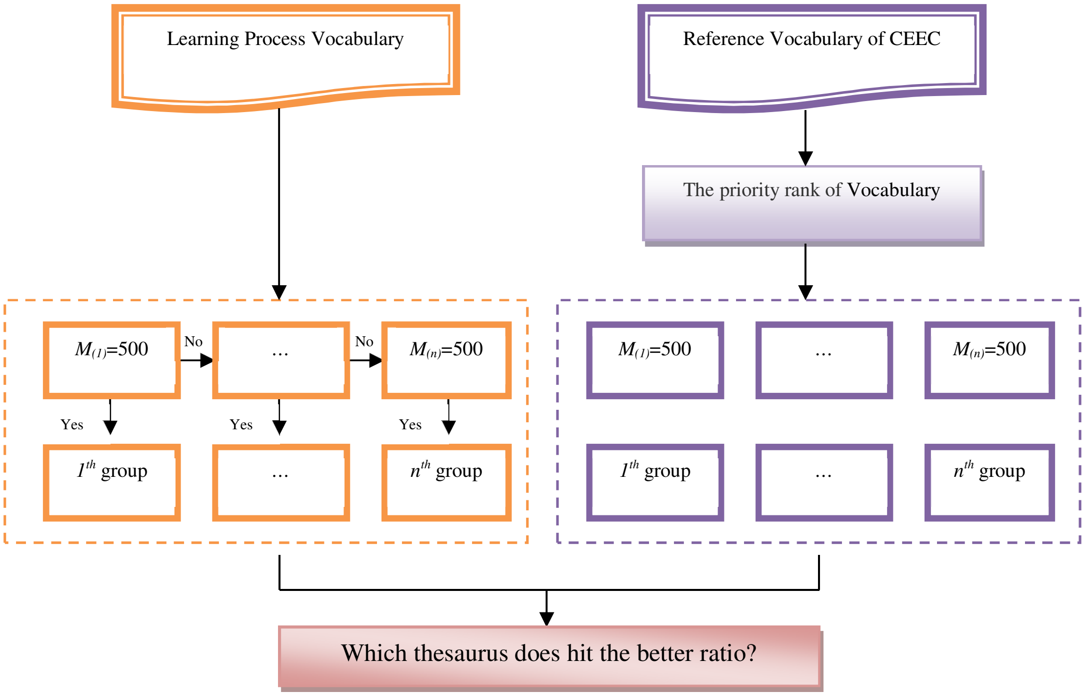

Learning Process Vocabulary
Reference Vocabulary of CEEC
The priority rank of Vocabulary
$M_{(1)}$=500
No
...
No
$M_{(n)}$=500
Yes
Yes
Yes
1th group
...
nth group
$M_{(1)}$=500
...
$M_{(n)}$=500
1th group
...
nth group
Which thesaurus does hit the better ratio?

Fig. 7. Sort the important vocabulary chart

A random sample was selected from the reference vocabulary provided by the CEEC. Each time, 500 words were randomly chosen to build an $M_{(i)}$ template (i = 1, 2,…, n), where *i* refers to a round of chosen words. The words were arranged into an $M_{(i)}$ group. The $M_{(j)}$ module is a product of VLPM construction. When the number of words in the vocabulary of the $M_{(j)}{}^{th}$ module equals 500, the words constitute the j^{th} vocabulary set; if the number of words in the vocabulary of the $M_{(j)}{}^{th}$ module is less than 500, the VLPM orders the vocabulary in the $M_{(j+1)}{}^{th}$ module alphabetically and implements random sampling to supplement the lack of words. The resulting vocabulary set is also denoted as the j^{th} vocabulary set. The rest can be deduced by analogy.

The process had the following steps. First, the number of words in every vocabulary list was counted. Second, a model for predicting words on future examinations was constructed. The results can provide a reference vocabulary for examination preparation.

5. Experimental Results and Analysis

The PCEECT, TbT, E&J, and CEECT were used as data sources. Table 8 shows the word counts of these sources.

1. The PCEECT contained vocabularies of both the SAT and DRT each year from the 2002 to 2011 academic years.
2. The TbT contained vocabulary from the textbooks of three major publishers, Lung Teng, San Min, and Far East, which are used by most senior high school students in Taiwan (Luo, 2010; Wang, 2006).
3. The E&J contained the vocabulary for elementary and junior high schools that the Ministry of Education announced as suggested learning. This source represents what students are expected to know in elementary and junior high school.

4. The CEECT is an approximately 7,000-word list recommended for senior high school students. Zheng (2002) indicated that the vocabulary list is referenced in teaching English and preparation courses for the SAT and DRT, but the range of English words in the college entrance examination is not limited to the list.

Table 8. The number of vocabulary in each thesaurus

Thesaurus	Past Examinations Thesaurus (PE)	Textbooks Thesaurus (TbT)			Basic 1000 Thesaurus (BT)	CEEC Suggested Thesaurus (CEECT)
		Lung Teng publisher	San Min publisher	Far East publisher		
The number of vocabulary	4,712	2,213	2,155	2,000	1,076	6,311

Section 5 is organized as follows. Section 5.1 presents a comparison of the suitability of the PCEECT with that of the CEECT. Section 5.2 explores the tacit patterns and correlations between the PCEECT and a target examination. Section 5.3 discusses the application of the VLPM in predicting words on future examinations. Section 5.4 details the prioritization of vocabulary in the VLPM for future examinations

.

5.1. Suitability check module for comparing the suitability of the PCEECT and the CEECT

The suitability of the PCEECT was examined and compared with that of the CEECT by using the suitability check module. The F-measure values of each year, shown in Fig. 8, reveal that the applicability of the PCEECT was superior to that of the CEECT regardless of the year. Hence, the PCEECT should be considered a study aid superior to the CEECT.

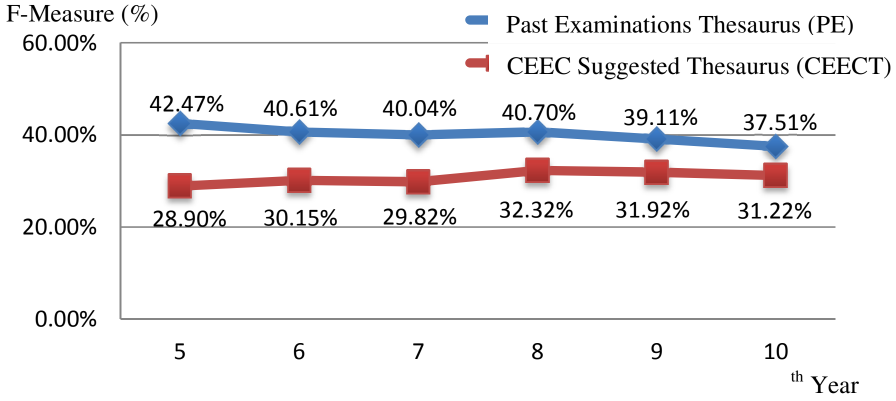


Fig. 8. Comparison with F-measure of PCEECT and CEECT in each year

The CEECT contained only 3,616 words out of the 6,311 total that had appeared in exams, representing more than 50% of the words used over the past decade. Therefore, the CEECT may not be suitable for an examinee who intends to strategically memorize and prepare a vocabulary list in a short time, especially in the final period of exam preparation. The word distributions of vocabulary lists are shown in Table 9. For example, the second row of the table indicates that 195 words appeared 10 times in examinations over the past decade. Among these words, 176 were included in the E&J and 17 were included in the TbT, and 2 words were not included in these two word lists.

Table 9. The frequency of past examinations during past ten years

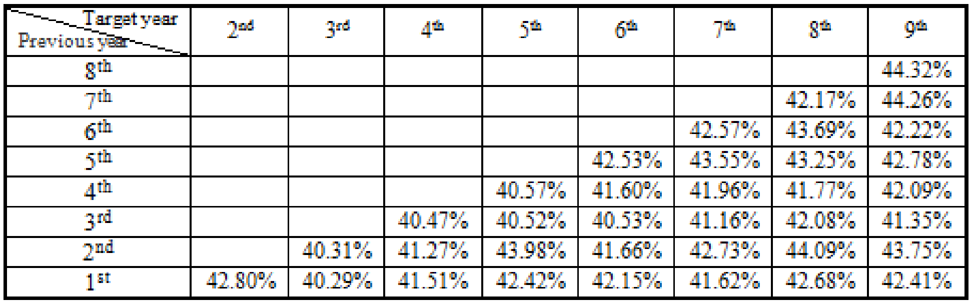
The frequency appear in past exams	Thesaurus of elementary school, and junior high school (E&J)	Thesaurus of textbooks (TbT)	Other Thesaurus	Total
10	176	17	2	195
9	58	31	3	92
8	74	39	1	114
7	66	64	4	134
6	64	100	5	169
5	62	117	8	187
4	78	193	18	289
3	89	302	62	453
2	85	449	145	679
1	122	739	443	1,304
0	**98**	**1,109**	**1,488**	**2,695**
total	**972**	**3,160**	**2,179**	**6,311**

Note: Other Thesaurus indicates the thesaurus which the vocabulary doesn't appear in E&J and TbT.

5.2. Tacit patterns and correlations between the PCEECT and the target examinations

This study explored the relationship between the target examinations and the past examinations. Because the data from target years were used as the model testing data sets, each examination's content in the year before the target year was considered the input training data set. In other words, when the second to ninth years of the CEEC examination were used as the target years, the past examinations from the first to eighth years were the input data used to calculate the content cosine similarity between the target year and the past year. The experimental results are provided in Table 10.

Table 10. Compare cosine similarity of past examination and target examination



Target year Previous year	2nd	3rd	4th	5th	6th	7th	8th	9th
8th								44.32%
7th							42.17%	44.26%
6th						42.57%	43.69%	42.22%
5th					42.53%	43.55%	43.25%	42.78%
4th				40.57%	41.60%	41.96%	41.77%	42.09%
3rd			40.47%	40.52%	40.53%	41.16%	42.08%	41.35%
2nd		40.31%	41.27%	43.98%	41.66%	42.73%	44.09%	43.75%
1st	42.80%	40.29%	41.51%	42.42%	42.15%	41.62%	42.68%	42.41%

The examinations in each year were ordered by decreasing value of cosine similarity and the three highest values were then selected. For example, in Table 11, the results indicate that the target examination for each year (fourth to ninth examinations) and second-year examination were the most similar among the first three past examinations. Therefore, the examinations in the first three years were critical. Furthermore, the second-year past examination content appeared in the first three past examinations. Thus, the second past examination was also critical. The ranks are illustrated in Table 11.

Table 11. The higher correlation for first three examinations

Target year / Previous year	4th	5th	6th	7th	8th	9th
8th						❶✓
7th						❷✓
6th				❸✓	❷✓	
5th			❶✓	❶✓	❸✓	
4th		❸✓				
3rd	❸✓					
2nd	❷✓	❶✓	❸✓	❷✓	❶✓	❸✓
1st	❶✓	❷✓	❷✓			

Note: the number in the cell represents the cosine similarity rank of the target year

According to the previous experimental results, the target vocabulary for the first three and the second-year past examinations is called the 3 + 2 vocabulary list in this paper. The results indicated that the reference value of the 3 + 2 vocabulary list was high, and the probability for the vocabulary list to match the content of the target year was approximately 45%. Hence, the examinees should consider the 3 + 2 vocabulary list when preparing for their examination.

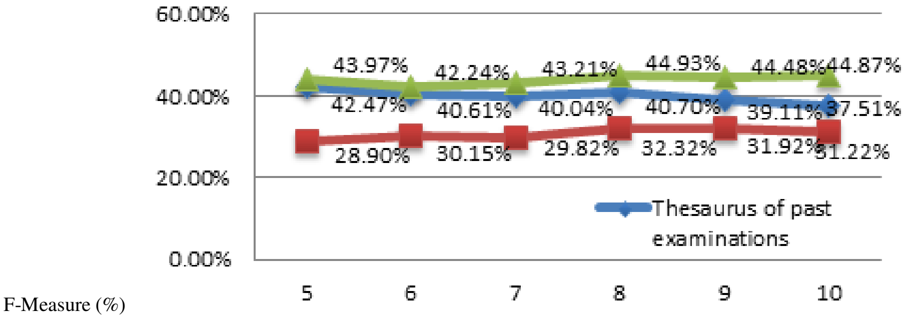


Fig. 9. The F-Measure of thesaurus of past examinations(PE), CEEC, 3+2

5.3. Application of the VLPM to future examinations

The study established an eighth-year VLPM, which used the E&J, TbT, and PCEECT from the first to seventh years as well as target vocabulary for the eighth year. Based on the relationships among these word lists, the VLPM was used to determine the optimal predictive model for the target year examination. The ninth-year target vocabulary was used as the testing data and the performance of the proposed model for the 10th year was verified.

The properties of each word can determine whether it is included in a vocabulary list. This study designed eight modules based on the E&J, TbT, and PCEECT to determine whether a word was included. The vocabulary list T indicates the result of the target year. Module 1 E&J(0)-TbT(0)-PCEECT(0) contained the words that were not included in the E&J, TbT, and PCEECT, but appeared in the eighth target year, shown as T(8), and matched 2.95% of the examination content. Furthermore, there are no words in T(8) that were not included in BT(0)-TbT(0)-PCEECT(0) because the joint probability was 0.

Year

This study ranked each module by the frequency at which each module appeared in the T(8) and T(9) vocabulary list. The experimental results in Table 12 and 13 show that the first four modules, E&J(0)-TbT(0)-PCEECT(0), E&J(1)-TbT(1)-PCEECT(1), E&J(1)-TbT(0)-PCEECT(1), and E&J(0)-TbT(1)-PCEECT(1) are the same in the eighth VLPM and the ninth

VLPM. The recall index of the first four modules in the eighth VLPM was 68.93%, and that in the ninth VLPM was 78.40%. Therefore, the vocabulary of the first four modules appeared frequently in the target examination.

Table 12. The ratio and cumulative number in the 8th target thesaurus

module	Thesaurus				Joint probability	The ratio of each module appear in $T_{(8)}$ Thesaurus	The cumulative number of vocabulary	The cumulative number of 9th vocabulary (W)	The ratio of 9th vocabulary which appear in examination [(W)÷1558]
	E&J	Tb	PE	$T_{(8)}$					
1	0	0	0	0	0.00%	100%	0	0	0
	0	0	0	1	2.95%				
2	1	1	1	0	0.83%	60.96%	148	95	6.10%
	1	1	1	1	1.30%				
3	1	0	1	0	4.42%	58.66%	896	523	33.57%
	1	0	1	1	6.28%				
4	0	1	1	0	19.99%	28.87%	2,928	1,074	68.93%
	0	1	1	1	8.11%				
5	1	1	0	0	0.18%	14.29%	2,940	1,077	69.13%
	1	1	0	1	0.03%				
6	1	0	0	0	2.45%	8.20%	3,108	1,090	69.96%
	1	0	0	1	0.22%				
7	0	0	1	0	21.21%	8.15%	4,892	1,242	79.72%
	0	0	1	1	1.88%				
8	0	1	0	0	28.60%	5.13%	6,852	1,351	86.71%
	0	1	0	1	1.55%				
Total					100%				

Table 13. The ratio and cumulative number in the 9th target thesaurus

module	Thesaurus				Joint probability	The ratio of each module appear in $T_{(9)}$ Thesaurus	The cumulative number of vocabulary	The cumulative number of 10th vocabulary (W)	The ratio of 10th vocabulary which appear in examination [(W)÷1338]
	E&J	Tb	E	$T_{(9)}$					
1	0	0	0	0	0.00%	100%	0	0	0
	0	0	0	1	2.93%				
2	1	1	1	0	0.14%	93.26%	151	105	7.85%
	1	1	1	1	1.96%				
3	1	0	1	0	1.40%	86.77%	912	486	36.32%
	1	0	1	1	9.20%				
4	0	1	1	0	8.88%	69.15%	3,053	1,049	78.40%
	0	1	1	1	19.91%				
5	0	0	1	0	13.17%	47.87%	5,044	1,197	89.46%
	0	0	1	1	12.10%				
6	1	1	0	0	0.13%	25.00%	5,053	1,199	89.61%
	1	1	0	1	0.04%				
7	1	0	0	0	2.20%	7.74%	5,208	1,211	90.51%
	1	0	0	1	0.18%				
8	0	1	0	0	26.22%	5.56%	7,059	1,303	97.38%
	0	1	0	1	1.54%				
Total					100%				

The joint probability for the module E&J(0)-TbT(0)-PCEECT(0) was 100%. This indicated that every year must have had some words not included in these vocabulary lists but that appeared in the target examination; however, because the number of vocabulary words in this category was quite high, the E&J(0)-TbT(0)-PCEECT(0) module was eliminated. Consequently, the model computes performance indices for only the second module (E&J(1)-TbT(1)-PCEECT(1)), third module (E&J(1)-TbT(0)-PCEECT(1)), and fourth module (E&J(0)-TbT(1)-PCEECT(1)) for the eighth VLPM and ninth VLPM. The F-measure was 47.88% for the examination in the ninth year and 47.78% for the examination in the 10th year. Therefore, examinees can prioritize learning the vocabulary words in the E&J(1)-TbT(1)-PCEECT(1), E&J(1)-TbT(0)-PCEECT(1), and E&J(0)-TbT(1)-PCEECT(1) modules, with approximately 48% of the modules' vocabularies appearing in the target examination, as shown in Table 14.

Table 14. The precision, recall, and F-Measure in the 9th and 10th target examination

Year	Thesaurus			The cumulative number of vocabulary	The cumulative number of vocabulary which appear in examination	Precision	Recall	F-Measure
	E&J	Tb	E					
9th	1	1	1	2,928	1,074	36.68 %	68.93 %	47.88 %
	1	0	1					
	0	1	1					
10th	1	1	1	3,053	1,049	34.36 %	78.40 %	47.78 %
	1	0	1					
	0	1	1					

5.4. Prioritized vocabulary in the VLPM for future examinations

In this study, an experiment was conducted to determine whether the proposed VLPM vocabulary list or the CEECT had a higher hit ratio for the target examination. Because the VLPM was constructed using the ranked modules of past years and different modules possess different volumes of vocabulary, for both the 9th and 10th examinations, the proposed VLPM vocabulary list, with the increase of vocabulary, had a higher hit ratio than that of the CEECT. This result suggests that examinees should focus on reviewing the proposed VLPM vocabulary list in preparation for their college entrance examinations, as shown in Table 15 and Table 16.

Table 15. The ratio of CEEC thesaurus, VLPM, for the 9th examinations

The number of vocabulary	CEEC thesaurus and 9th examination			Thesaurus of Vocabulary Learning Process Model and 9th examination		
	Intersection number of vocabulary	The cumulative number of intersection vocabulary	**Hit-ratio**	Intersection number of vocabulary	The cumulative number of intersection vocabulary	**Hit-ratio**
1-500	273	273	**18%**	287	287	**18%**
501-1000	338	611	**39%**	261	548	**35%**
1001-1500	81	692	**44%**	139	687	**44%**
1501-2000	128	820	**53%**	136	823	**53%**
2001-2500	38	858	**55%**	146	969	**62%**
2501-2928	119	977	**63%**	105	1,074	**69%**

Finally, E&J$_{(1)}$-TbT$_{(1)}$-PCEECT$_{(1)}$, E&J$_{(1)}$-TbT$_{(0)}$-PCEECT$_{(1)}$, and E&J$_{(0)}$-TbT$_{(1)}$-PCEECT$_{(1)}$ were used to predict the words on the 11th-year English examination. A total of 3,159 words were deemed essential and showed a high hit ratio. This indicates how examinees can prioritize their review of vocabulary words in a limited time, as shown in Table 17.

Table 16. The ratio of CEEC thesaurus, VLPM, for the 10th examinations

The number of vocabulary	CEEC thesaurus and 10th examination			Thesaurus of Learning Process Vocabulary and 10th examination		
	Intersection number of vocabulary	The cumulative number of intersection vocabulary	**Intersection ratio**	Intersection number of vocabulary	The cumulative number of intersection vocabulary	**Intersection ratio**
1-500	398	398	**30%**	287	287	**21%**
501-1000	243	641	**48%**	246	533	**40%**
1001-1500	18	659	**49%**	269	802	**60%**
1501-2000	115	774	**58%**	86	888	**66%**
2001-2500	32	806	**60%**	92	980	**73%**
2501-3000	118	924	**69%**	66	1,046	**78%**
3001-3053	4	928	**69%**	3	1049	**78%**

Table 17. The important term list of English examination in 101 college entrance

Module	Suggestive rank	The number of vocabulary	The cumulative number of vocabulary
E&J$_{(1)}$-Tb$_{(1)}$-E$_{(1)}$	1	153	153
E&J$_{(1)}$-Tb$_{(0)}$-E$_{(1)}$	2	773	926
E&J$_{(0)}$-Tb$_{(1)}$-E$_{(1)}$	3	2,233	3,159
total		3,159	

6. CONCLUSIONS

In this study, vocabulary lists, namely the E&J and TbT, were combined with past exam items to develop the VLPM. This study determined that the words on past examinations are relevant, evidencing that studying the vocabulary words on past exam items can help an examinee prepare for future examinations. Furthermore, this study identified the regularities and relationships between the target and past examinations. The content of the second-year CEEC examination seemed to be a paradigm for the other examinations. Its content had a relatively high cosine similarity with that of the other examinations. Finally, this study provides a prioritized vocabulary list based on the examination strategy. The purpose of the VLPM is to help examinees study and receive high examination scores in a limited time.

The proposed method can also apply in other language learning, especially in examination strategy. For example, while in German vocabulary learning, the users just have to replace the examination content as German official test items. Besides, finding the different thesaurus features to identify different vocabulary. Exploiting the machine learning steps to discover the patterns which between the thesaurus and examinations. Finally to get the different language's learning list.

ACKNOWLEDGEMENT

We thank the editor and anonymous reviewers for their constructive comments, which have greatly improved the readability and quality of this article. This research is supported by Fu Jen Catholic University A0103040.

REFERENCES

[1] AbuSeileek, A. F. (2011). Hypermedia annotation presentation: The effect of location and type on the EFL learners' achievement in reading comprehension and vocabulary acquisition. Computers & Education, 57, 1281-1291.

[2] Astika, G. G. (1993). Analytical Assessments of Foreign Students' Writing. RELC Journal, 24, 61-70.

[3] Boonchom, V., & Soonthornphisaj, N. (2012). ATOB algorithm: an automatic ontology construction for Thai legal sentences retrieval. Journal of Information Science, 38, 37-51.

[4] Chen, C. M., & Chung, C. J. (2008). Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle. Computers & Education, 51, 624-645.

[5] Cheng, H. H. (2002). Reference Words of High school English. Taipei: Center of College Entrance Examination.

[6] Chou, S. Y. (2009). A Study of Cloze Test Items in Scholastic Aptitude English Test and Department Required English Test. National Chung Cheng University, Department of Foreign Languages and Literature master thesis.

[7] Fang, Y. S. (2008). A Comparison of Scholastic Aptitude English Test and Department Required English Test. National Tsing Hua University, Department of Foreign Languages and Literature master thesis.

[8] Hinkel, E.(2006). Current Perspectives on Teaching the Four Skills. Tesql Quarterly, 40, 109-131.

[9] Hsu, C. K., Hwang, G. J., & Chang, C. K. (2013). A personalized recommendation-based mobile learning approach to improving the reading performance of EFL students. Computers & Education, 63, 327-336.

[10] Hsu, M. H. (2008). A personalized English learning recommender system for ESL studetns, Expert Systems with Applications, 34, 683-688.

[11] Huang, Y. M., Huang, Y. M., Huang, S. H., Lin, Y. T. (2012). A ubiquitous English vocabulary learning system: evidence of active/passive attitudes vs. usefulness/ease-of-use. Computers & Education, 58, 273-282.

[12] Jia, J., Chen, Y., & Ding, Z. (2012). Effects of a vocabulary acquisition and assessment system on students' performance in a blended learning class for English subject. Computers & Education, 58, 63-76.

[13] Laufer, B., & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. Applied Linguistics, 16, 307-322.

[14] Li, C.H., Yang, J. C., & Park, S. C. (2012). Text categorization algorithms using semantic approaches, corpus-based thesaurus and Word Net. Expert Systems with Applications, 39, 765-772.

[15] Lin, M. C. (2007). Effects of online academic vocabulary instruction on EFL college writing. National Tsing Hua University, Department of Foreign Languages and Literature master thesis.

[16] Luo, C., Li, Y., & Chung, S. (2009). Text document clustering based on neighbors. IEEE Transaction on Data & Knowledge Engineering, 68, 1271-1288.

[17] Osman, A. H., Salim, N., & Binwahlan, M. S. (2012). An improved plagiarism detection scheme based on semantic role labeling. Applied Soft Computing, 12, 1493-1502.

[18] Sandberg, J. Maris, M. & Hoogendoorn, P. (2014). The added value of a gaming context and intelligent adaptation for a mobile learning application for vocabulary learning. Computers & Education, 76, 119-130.

[19] Schmitt, N. (2000). Vocabulary in language teaching. Cambridge University Press.

[20] Smith, G. G., Li, M., Drobisz, J., Park, H. R., Kim, D., & Smith, S. D. (2013). Play games or study? Computer games in eBooks to learn English vocabulary. Computers & Education, 69, 274-286.

[21] Uguz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems, 24, 1024-1032.

[22] Wu, Q. (2015). Designing a smartphone app to teach English (L2) vocabulary. Computers & Education, 85, 170-179.

[23] Yang, I. L. (2006). On the Issue of Vocabulary Size in English Teaching in Taiwan. Journal of the National Institute for Compilation and Translation, 34, 35-44.

[24] Zheng, H., Chen, J. Y., & Jiang, Y. (2012). An ontology-based approach to Chinese semantic advertising. Information Sciences, 216, 138-154.