Development of Word-Level Classification and Vocabulary Meaning System

Kamal Baha and Makoto Shishido

***Abstract*—There are multiple ways to improve reading comprehension for English learners as a foreign language. Learning vocabulary is one of the ways to improve it. It is believed that the more readers are familiar with the English vocabulary, the better they will understand what they are reading. It is suggested that the learners improve comprehension if they learn and understand unknown words before they read an essay. The morphological analysis was used to extract the words from each sentence in an English text. The extracted vocabulary was also classified by level of difficulty into 12 levels according to the ALC12000 database. The system showed only the words whose levels were higher than the student's estimated vocabulary level. It will be expected that a learner can improve their English reading comprehension by studying higher-level vocabulary in advance of reading an essay. In this study, a new method was investigated in order to analyze vocabulary and develop a system that can analyze vocabulary from English texts and add Japanese and Thai meanings. To employ it for any device, the researchers developed the system by using a JavaScript library called NLP-compromise that uses any browser on any system.**

***Index Terms*—Educational technology, computer aided instruction, system development, language processing, morphology, word-level classification.**

I. Introduction

English is the most used international language globally, so learning English has become essential. Students need to learn more vocabulary to understand the meanings of English sentences and texts [1], [2]. Nevertheless, many students think learning English vocabulary is complicated, especially for words they do not know or do not use frequently. A learner must know at least 2,000 commonly used English words [3], [4] to read English texts efficiently and fluently. The acquisition of new vocabulary is crucial for English study, yet classroom time is limited. Additionally, learning English includes different skills, such as listening, reading, writing, and speaking. Therefore, teachers should consider strategies to help students acquire English vocabulary outside the classroom.

This study aimed to analyze English words or vocabulary from any English text using various methods. Vocabulary learning can be difficult for English beginning learners [5], [6]. It is important to support English beginning learners to acquire unknown words or vocabulary in advance before they start reading a text since it is believed that the more readers know the English vocabulary, the more the readers understand what they are reading [7], [8].

The simplest method for analyzing words is to have humans with superior English language abilities read and separate the vocabulary from the text. Besides, it still needs to manipulate the difficulty levels and compare the vocabulary difficulty levels with a vocabulary database. Nevertheless, this method is unsuitable for people who are required to read and separate vocabulary but cannot do it in large quantities. It also results in a high rate of inaccuracy.

In this study, a new method is proposed to analyze English vocabulary from any English text using a computer program. Furthermore, the researchers developed a system for analyzing English words or vocabulary from any English text, compare the difficulty with the vocabulary database, and automatically add vocabulary's meanings in Thai and Japanese. Thus, this study aims to develop an English vocabulary learning support system for the senior high school students in Thailand and Japan.

We focused on comparing the behavior performance of students to the systems of the two countries. Since we used the English-Japanese vocabulary database (ALC12000), Thailand has not yet developed a system to extract words from any English text that are automatically classified by level and added Thai meaning. In addition, both countries utilize English as a foreign language. We will compare the various perspectives on English vocabulary acquisition and how to develop an appropriate system for vocabulary learning in each country.

II. Literature Review

A. Waterfall Methodology

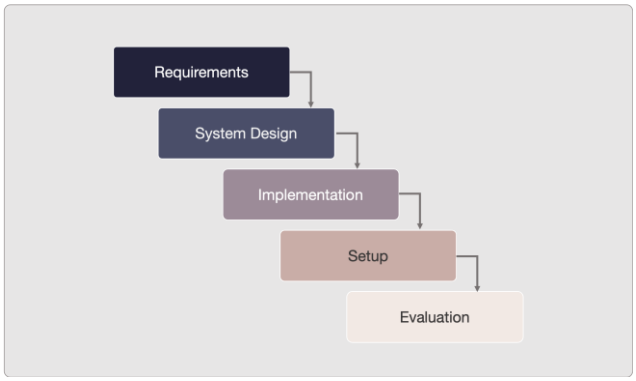

Fig. 1. The waterfall methodology model.

The Waterfall method, a systematic and logical model for software development systems [9], was used as the system development methodology in this study. The Waterfall or Linear Sequence model is frequently referred to as the

Manuscript received April 14, 2022; revised June 17, 2022.

Kamal Baha and Makoto Shishido are with Tokyo Denki University, Japan (e-mail: 20udc02@ms.dendai.ac.jp).

doi: 10.18178/ijiet.2022.12.11.1740

"Classic Life Cycle" [10]. The essence of the Waterfall method is that work on a system is performed sequentially because each stage must wait for the completion of the previous one. the Waterfall software development methodology, which consists of five stages: requirements, system design, implementation, setup, and evaluation (Fig. 1).

B. Constructivism Theory

The main principle of the educational philosophy known as Constructivism is that learners should be taught to be self-sufficient and responsible for their own learning [11]. Constructivism describes to how of learning and thinking. The term "Constructivist" refers to a teaching method that emphasizes the students' active participation in the learning process and how content can be efficiently communicated to students. [12], Constructivism believes that learners' perceptions of knowledge are generated from a meaning-making exploration in which learners develop individual interpretations of their experiences [13].

III. Related Works

Koichi Higuchi [14], [15] explained that the KH Coder was a free program for conducting quantitative content analysis and text mining. Additionally, it was used in computational linguistics. However, KH Coder lacked many English vocabulary analysis functions, such as the inability to level the difficulty of vocabulary due to the absence of an initial database of vocabulary level difficulty. It was also incapable of translating the words it analyzes into other languages.

New Word Level Checker (NWLC) [16] was a web app for vocabulary profiling and analyzes English words submitted by the user and produces vocabulary levels based on the selected word lists. The NWLC system was developed and programmed by Dr. Atsushi Mizumoto which utilization of the concept of The Word Level Checker [17]. However, the major challenge of both systems is their inability to show words as text data. As a result, the system could not analyze some text data and include the vocabulary meaning in other languages, such as Thai and Japanese.

Quizlet was a unique online vocabulary management system that could run on a mobile device/laptop as an application and a website and provided users with seven practical vocabulary learning tools for constructing a variety of vocabulary exercises. Nevertheless, the Quizlet could not analyze English vocabulary from English texts. Instead, it needs the user to input the vocabulary and meaning of the word into the Quizlet system itself. This method makes it very inconvenient when many words need to be entered [18].

IV. Study

In the requirements stage, we analyzed the lack of features in previous works, including KH coder [14], New Word Level Checker (NWLC) [16], and Quizlet, to propose a new system with the new features.

We believed that learning vocabulary in advance before reading English text was able to help the learners understand English text. The system could morphologically extract words from any English text and transform grammatically changed words' forms into base forms, e.g., plural nouns or past tense forms. The extracted words were classified level difficulty as SVL12000 standard vocabulary level database, containing 12000 English words and divided into 12 difficulty levels. The system could automatically add English words meaning in Thai (English–Thai Cambridge Dictionary) and Japanese (ALC Education Inc.). Furthermore, we developed the system by using website development technologies. The learners(users) could access this learning system even with mobile, tablet devices, or laptops.

It was hoped that the developed Word-Level Classification and Vocabulary Meaning system would be related to the innovative learning system and contribute to the new educational technology to support English beginning learners (Thai and Japanese) to acquire unknown words or vocabulary in advance before they start reading text and improve their English reading comprehension.

The researchers aimed to define the analyzed word and name this new system as the word level classification and vocabulary meaning system. In the first stage of system development, the researchers used various tools. As a result, it was difficult for use. Then, in the second stage, the researchers tried to make it easier to use by developing a new system to simplify the implementation process and be able to use it online.

A. Word Level Classification and Vocabulary Meaning System

Word level classification and vocabulary meaning system is a system that can analyze English words or vocabulary from any English essay. It can also identify learner's vocabulary levels. It can automatically add vocabulary meaning based on the vocabulary level's difficulty, allowing learners to analyze vocabulary in English texts and study vocabulary before reading it. The learners will study unknown words in the English essay in advance, and the content will become easier to understand when they read an English essay.

B. Text Analysis Tools

1) KH coder

KH Coder is an open-source software for statistical analysis of text data [14], [15]. It was created to analyze various social survey data, such as free descriptions of questionnaires, interview records, and newspaper articles. It is compatible with methods called "metric text analysis" or "text mining."

2) Features

KH Coder is produced assuming two stages of the analysis procedure. In the first stage, words are automatically extracted from the data, and the results are tabulated and analyzed. It allows us to explore the characteristics of the data and summarize the data as possible. In the second stage, the analyst actively and explicitly specifies (if there is such an expression, it is assumed that concept A has appeared) (creating coding rules) and extracts the concept from the data. Aggregate and analyze the results to deepen the analysis.

3) What KH coder cannot do?

KH Coder software cannot level vocabulary with other vocabulary databases. Moreover, only using it offline also runs on the Windows 10 for free. Therefore, it cannot be used in a smartphone or tablet (Online).

C. Solutions

For the use of KH Coder in developing vocabulary analysis systems, there are limitations in terms of equipment, operating systems, and that users are unable to use the system on all devices. The researchers created a software with an online platform (Website) to analyze English essays. Users can easily use the system by copying a text and pasting on our vocabulary level checking system.

V. System Development and Design

The system development was divided into two phases. In the first phase, the researchers used the developed program, such as KH Coder, to analyze essays and reduce the entire system's development time. It can immediately be able to apply the analyzed vocabulary data. In the second part, the researchers used web development technology in our system development to reduce the complexity of using the system.

A. Conventional Tools

Word level classification and vocabulary meaning system is a system that can analyze English words or vocabulary from any English essay. It can also identify learner's vocabulary levels. It can automatically add vocabulary meaning based on the vocabulary level's difficulty, allowing learners to analyze vocabulary in English texts and study vocabulary before reading it. The learners will study unknown words in the English essay in advance, and the content will become easier to understand when they read an English essay.

1) Tools for using in our system

Many tools were used to create a level checking system. The First is KH Coder, which is an open-source program for analyzing English essays. The program will show that the results are vocabulary, parts of speech, and frequency in the order of frequency number (Fig. 2). The researchers exported the analyzed vocabulary data to an Excel file for categorizing with our program.

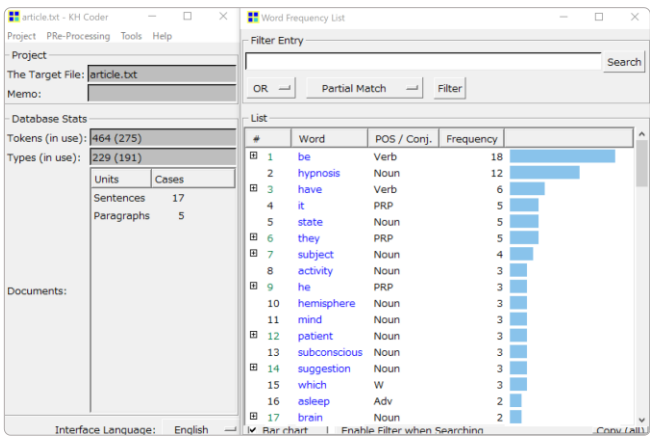

Fig. 2. The analyzed vocabularies data display their part of speech and frequency.

The second is the ALC vocabulary database (SLV12000) [19] that is the vocabulary database from ARUKU company, and it is divided into 12 difficulty levels. The next one is the Java program (Fig. 3) that the researchers developed to level vocabulary level and add Japanese meaning. The last one is Quizlet, which is a vocabulary study website (Fig. 4).

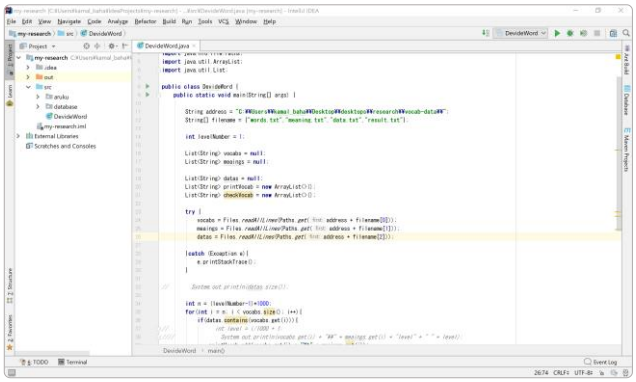

Fig. 3. The Java program that we developed for leveling vocabulary level and adding Japanese meaning.

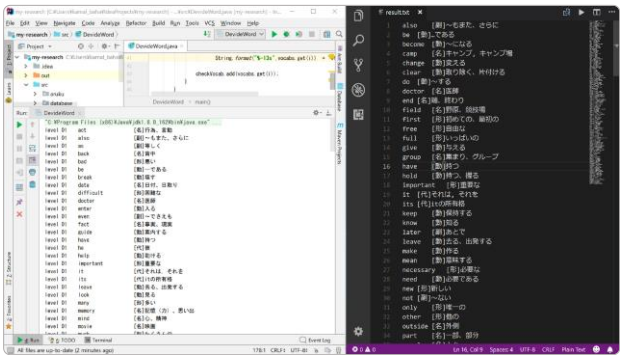

Fig. 4. The added meaning text file for the vocabulary learning material by Quizlet.

2) Conventional method flow

A piece of English text was used to try out. Fig. 5 shows the conventional English text analysis method flow. Firstly, the text was analyzed using KH Coder. Secondly, the analyzed data by KH Coder were exported in an excel file format. The Java program then identifies the vocabulary level. Furthermore, the meaning of Japanese vocabulary will be added by the Java program. Finally, the Quizlet website will then create the vocabulary learning material.

Fig. 5. Conventional method flow.

B. Text Analysis Tools

The level checking system that has been conventionally developed is inconvenient for users or learners. It is because

the installation flow is complicated, and there are many tools and processes when setting it up and using it. These tools can only be used on Windows 10 operating system offline. In this study, the researchers developed a system for English essay analysis on a web platform to be accessed online and used in a simple process. It does not depend on only one OS or device, such as the previous version (Fig. 6).

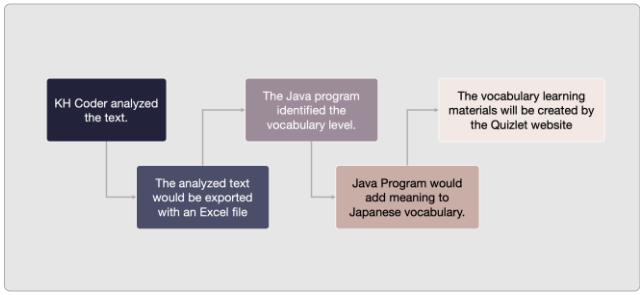

Fig. 6. The concept of one system in proposed system.

To analyze the English texts, the process was divided into two parts for analyzing, and different tools were developed to analyze each part

1) Software user interface

The researchers mainly developed our system using web development tools and programming languages, such as HTML, CSS, and JavaScript. The first page is the main page with a menu of various futures that we will be developed further in future work (Fig. 7).

Analyzing English essays will open when the user clicks on the get start button in the middle of the main page.

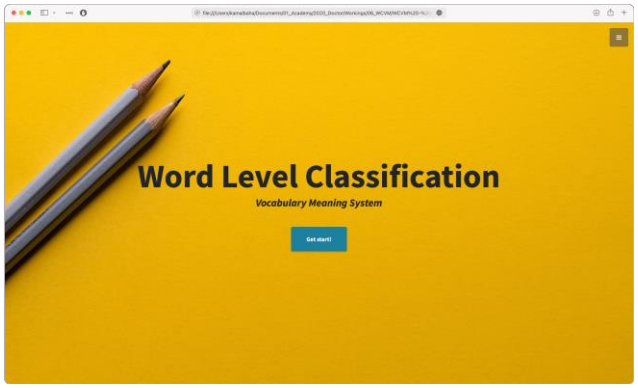

Fig. 7. The first page of system user-interface.

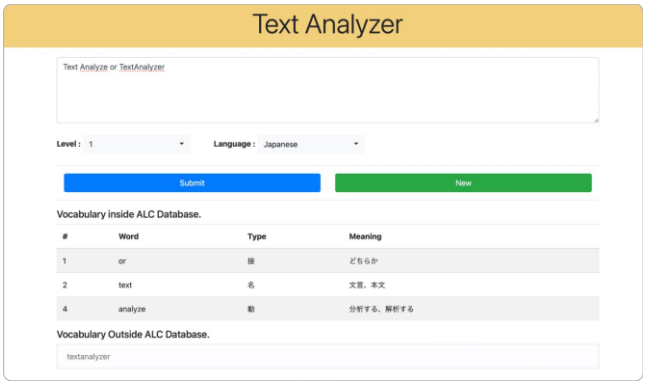

Fig. 8. The second page of system user-interface.

For the page used to analyze the English essays, the researchers designed it to be as easy to use as possible, with a text box providing the English text's content, which can put an extended essay. It is sufficient to analyze multiple page lengths of English essays without causing system problems.

In the next section, there will be a button. There are 12 levels of vocabulary difficulty to choose from. The level of difficulty of the words to be displayed depends on the level of the user chose. The system will display the vocabulary level from the user-selected level to the highest level available.

Next to the vocabulary difficulty button, there is a language menu to choose between two translation languages: Japanese and Thai (Fig. 8).

2) The system algorithm

In the first part, we use JavaScript programming to separate words in English sentences using space symbols. The system deleted the period symbol or a question mark from the separated words (Fig. 9).

```
//get text data from text area
let text = document.getElementById("myText").value.toString();

//analyse passage with nlp-compromise
let doc = nlp(text);

//seperate a word in passage with without check verb and noun form
let vocabs = doc.terms().data().map(x => x.text.replace(/[^\w ]/, ''));

//get only infinitive form of verb vocab in passage
let verbVocabs = doc.verbs().conjugate().map(x => x.Infinitive.replace(/[^\w ]/, ''));

//get only sigular form of noun vocab in passage
let nounVocabs = doc.nouns().data().map(x => x.singular.replace(/[^\w ]/, ''));
```

Fig. 9. The algorithm for separate words in sentences.

In the second part, the researchers tried to analyze the transformed words such as the plural word form of a noun that has been added s or es. Plural forms of nouns cannot be searched in the SLV12000 database and must be transformed into singular word forms. Similarly, the verb's past tense (-ed) and progressive tense (ing) need to be transformed into their original form or infinitive form. In this study, it was analyzed with a JavaScript library called NLP-compromise [20] and transformed into a form that can be searched in the SLV12000 database.

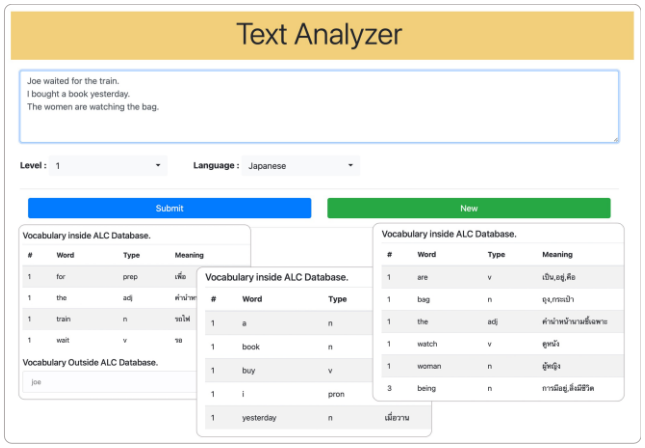

Fig. 10. The result display transformed words which analyze by our system.

Fig. 10 shows examples of English sentences analyzed by the system. The researchers present examples with various sentence patterns to try the analysis.

"Joe waited for the train." the system changed past simple tense "waited" to wait. and "I bought a book yesterday." it changed "bought" to "buy".

"The women are watching the bag." the system changed

the plural word form "women" to be "woman" and changed "watching" to be "watch".

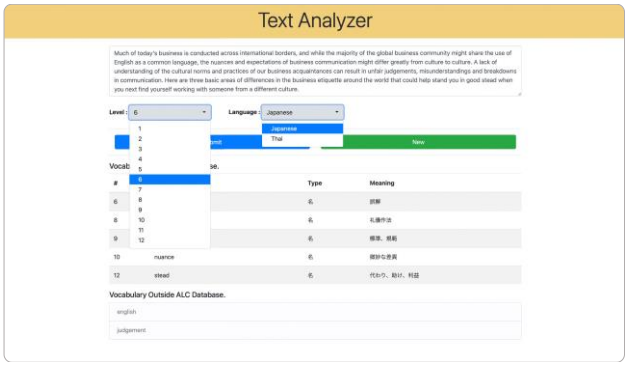

Fig. 11. The system can analyze even the extended essay.

Besides the 3 sentences that we had mentioned, the system is able to analyze long English essays (Fig. 11).

When the user selects the translation language that user wants it to display in Japanese, the system will display as in Japanese ((Fig. 12). Also, when selecting a Thai translation, it will display the selected Thai language ((Fig. 13).

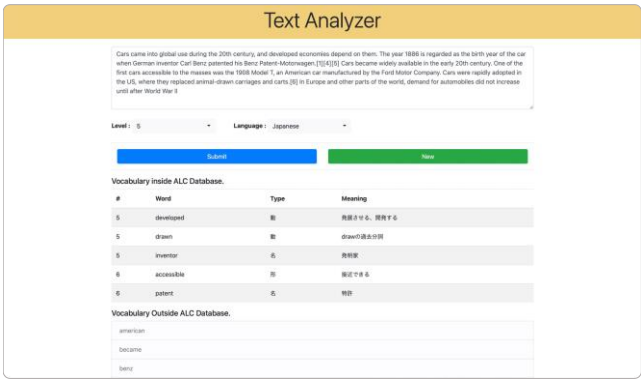

Fig. 12. The analyzed example essay and displays in Japanese.

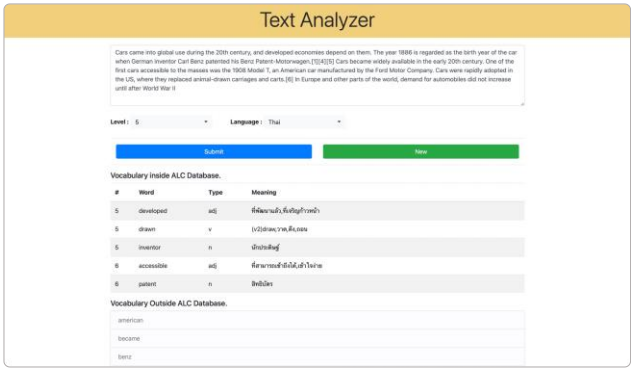

Fig. 13. The analyzed example essay and displays in Thai.

VI. Conclusion and Future Work

KH Coder system can only analyze English texts. However, it cannot level the vocabulary with other vocabulary databases. Moreover, only using offline also runs on Windows 10 OS for free. Therefore, it cannot be used by a smartphone or tablet (Online). The level checking system that has been developed conventionally is inconvenient for users or learners. It is because the installation flow is complicated, and there are many tools and processes when setting it up and using it.

In this study, the researchers developed a system for English essay analysis on a web platform. We used JavaScript programming to separate words in English sentences using space symbols. The system deleted the period symbol or a question mark from the separated words. Therefore, it can be accessed online and used in a simple process. It does not depend on only one operating system (OS) or device, such as the previous version.

The number of unknown vocabulary of users is different. Although the systems that we developed can analyze English texts and displays a higher level of vocabulary than the user's vocabulary-level, there may be some words that the user does not know on a lower difficulty level than the user's vocabulary-level.

Therefore, the researchers intend to develop new features that automatically create various types of vocabulary learning materials: Spelling, Dictation, Multiple Choice, Matching, True/False, and Flashcard games. Furthermore, the system will allow users to create a personal account in our future study, collect the history of system usage, and analyze the user's vocabulary level, including collecting unknown vocabulary and creating a database for the user individually. Moreover, we will evaluate the system's effectiveness by analyzing the results of the behavior performance of students to the systems compared between Thai and Japanese students by questionnaire.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

K. Baha developed the idea of research. Then, he coded the WCVM system used in this study. Finally, M. Shishido collected the data and helped in editing the paper.

References

[1] J. M. Harmon, "Teaching independent word learning strategies to struggling readers," *J. Adolesc. Adult Lit.*, vol. 45, no. 7, pp. 606–615, 2002.

[2] W. H. Rupley, J. W. Logan, and W. D. Nichols, "Vocabulary instruction in a balanced reading program," *Read. Teach.*, vol. 52, no. 4, pp. 336–346, 1998.

[3] D. D. Qian, "Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective," *Lang. Learn.*, vol. 52, no. 3, pp. 513–536, 2002.

[4] K. Mezynski, "Issues concerning the acquisition of knowledge: Effects of vocabulary training on reading comprehension," *Rev. Educ. Res.*, vol. 53, no. 2, pp. 253–279, 1983, doi: 10.3102/00346543053002253.

[5] H. H. Alghamdi, "Exploring second language vocabulary learning in ESL classes," *Engl. Lang. Teach.*, vol. 12, no. 1, Art. no. 1, Dec. 2018, doi: 10.5539/elt.v12n1p78.

[6] A. Linda and P. M. Shah, "Vocabulary acquisition style in the ESL classroom: A survey on the use of vocabulary learning strategies by the primary 3 learners," *Creat. Educ.*, vol. 11, no. 10, pp. 1973–1987, 2020, doi: 10.4236/ce.2020.1110144.

[7] D. H. Manihuruk, "The correlation between EFL students' vocabulary knowledge and reading comprehension: A case study at the English Education Department of Universitas Kristen Indonesia," *JET J. Engl. Teach.*, vol. 6, no. 1, pp. 86–95, 2020.

[8] D. Nunan, *Practical English Language Teaching*, 1st ed. New York: McGraw-Hill/Contemporary, 2003.

[9] D. R. Lucitasari, M. S. A. Khannan *et al.*, "Designing mobile alumni tracer study system using waterfall method: An Android based," *Int. J. Comput. Netw. Commun. Secur.*, vol. 7, no. 9, pp. 196–202, 2019.

[10] W. Suryn, *Software Quality Engineering: A Practitioner's Approach*, John Wiley & Sons, 2013.

[11] B. Patten, I. Arnedillo Sánchez, and B. Tangney, "Designing collaborative, constructionist and contextual applications for handheld devices," *Comput. Educ*., vol. 46, no. 3, pp. 294–308, Apr. 2006, doi: 10.1016/j.compedu.2005.11.011.
[12] N. H. Mvududu and J. Thiel-Burgess, "Constructivism in practice: The case for English language learners," *Int. J. Educ.*, vol. 4, no. 3, pp. p108–p118, Sep. 2012, doi: 10.5296/ije.v4i3.2223.
[13] H. Bauersfeld, "'Language games' in the mathematics classroom: Their func-tion and their effects," *The Emergence of Mathematical Meaning*, Routledge, 2012, pp. 277–297.
[14] K. Higuchi. (2016). KH coder 3 reference manual. Ritsumeikan University. [Online]. Available: https://khcoder.net/en/manual_en_v3.pdf
[15] K. Higuchi, "Using KH coder in the field of linguistics," *Math. Linguist. Soc. Jpn*., vol. 31, no. 1, pp. 36–45, 2017.
[16] A. Mizumoto. (2021). New word level checker [Web application]. [Online]. Available: https://nwlc.pythonanywhere.com
[17] S. Yasumasa. (2006). Aoyama gakuin university word level checker. [Online]. Available: http://someya-net.com/wlc/index.html?fbclid=IwAR3IbTZSGadkvMMyq5ngLXoBxKFpNa3Vfcq3RVNJ6DtrM-o6YQDMaVjd-fU
[18] M. S. Andarab, "Learning vocabulary through collocating on quizlet," *Univers. J. Educ. Res.*, vol. 7, no. 4, pp. 980–985, Apr. 2019, doi: 10.13189/ujer.2019.070409.
[19] Alc Education Inc. (2011). *Standard Vocabulary List SVL12000*. [Online]. Available: https://www.alc.co.jp/vocgram/article/svl/
[20] S. Kelly. (2018). Morphological analysis library. *Compromise*. [Online]. Available: http://compromise.cool/

Kamal Baha is a PhD student majoring in information, communication and media design engineering in Graduate School of Advanced Science and Technology at Tokyo Denki University. He was born in Thailand in 1993. His mother tongues are not only Thai but also the Malaysian language from his childhood. When he graduated from of Thailand to study in Japan. He seriously began high school in Thailand in 2012, he received a scholarship from the Ministry of Education to learn Japanese and English until he finished his bachelor's degree in the School of Information Environment at Tokyo Denki University in 2018. After completing his bachelor's degree, he was awarded the Ministry of Science scholarship of the Thai government again. He studied for his master's degree in the Graduate School of Information Environment at Tokyo Denki University and graduated in 2020. His main interest is to develop morphological analyzing systems for English vocabulary to help non-native speakers of English improve reading skills. His research interest includes programing and creating tools with modern technology to enhance foreign language learning.

Makoto Shishido is currently a professor of English and educational technology at School of System Design and Technology, Tokyo Denki University, Japan. He also serves as director of Center for International Affairs. After receiving his bachelor's degree in linguistics in Japan, Dr. Shishido attended University of the Pacific, Stockton, California in the U.S. gaining his master's degree in Education with an emphasis on multicultural education. He then started his Ph.D. study at International Christian University in Japan and earned his doctoral candidacy before embarking on an academic career, which has both embraced teaching undergraduate and postgraduate students and research. He then earned his Ph.D. degree at Tohoku University. Dr. Shishido's research has embraced studying the Input Hypothesis in second language acquisition and developing computer-assisted language learning materials for facilitating autonomous learning. His current research interests include developing e-learning materials with speech recognition and artificial intelligence and assessing their effectiveness.