AUTOMATIC SPEECH RECOGNITION SYSTEM TO ENHANCE THE USE OF VOCALIZATION STRATEGY

Saki Hirata[1] and Masanori Yamada[2]
[1]*Graduate School of Human-Environment Studies, Kyushu University*
[2]*Faculty of Arts and Science, Kyushu University*
744, Motooka, Nishi-ku, Fukuoka 819-0395, Japan

ABSTRACT

Vocabulary learning that overlooks the discrimination of single sounds in English language learning could lead to communication difficulties, such as the inability to hear and speak English pronunciation. To address this issue, vocalization strategy, in which students listen to the sounds of words and memorize them aloud, is considered a useful learning strategy, and has been reported to influence performance in English language learning. Many studies have shown that metacognition promotes the use of this strategy. In this study, we developed an English vocabulary learning support system using dictation-type speech recognition technology and evaluated whether it activates learners' metacognition and promotes the use of vocalization strategies.

KEYWORDS

Language Learning, Learning Strategies, Metacognition, Speech Recognition Technology

1. INTRODUCTION

1.1 Vocalization Strategy

Vocabulary is an important determinant of language proficiency. In second language learning (SLL), Shimamoto (2002) reported a significant correlation between learners' vocabulary levels and their overall TOEIC score. In vocabulary learning, memorizing the phonetic sounds of words is also considered an important learning component, and Nation (2001) mentions the multifaceted nature of vocabulary proficiency and considers knowledge of phonetic information of words as an aspect of vocabulary proficiency. It has also been reported that strengthening the association between written and speech forms can improve word listening (Field 2004) and appropriate speech (Li and Woore 2021).

Vocabulary learning strategies used by EFL learners, such as Japanese learners of English, have been reported to use bilingual dictionaries or vocabulary books to match the letters of words with their translations (Fan 2003; Schmitt 1997). In word learning through letters, "negative transfer," errors that arise from applying the rules of the native language to a foreign language, can result in learning incorrect pronunciations from letters (Odlin 1989). Negative transfer suggests that vocabulary learning that ignores the discrimination of single sounds can cause communication problems, such as the inability to hear and speak English pronunciation (Teshima 2011). Japanese learners have been reported to face difficulties recognizing and pronouncing English sounds that are not present in their native language, such as confusion regarding the difference between “r” and “l” phonemes in English (Ota et al. 2009).

While many studies have explored the effectiveness of explicit pronunciation instruction (e.g., Huo and Wang 2017), it has been reported that the success of vocabulary acquisition is determined by learners' individual learning (Rasouli and Jafari 2016). One of the factors determining the effectiveness of independent learning is learning strategy, which has been addressed in previous studies. As a strategy related to the phonetic information of words, Oxford (1990) suggested a pronunciation strategy in which students memorize words while pronouncing them. Conversely, it has been pointed out that in the wake of technological development, learning with audio equipment is gaining increasing popularity. As a strategy that reflects this change in the

foreign language learning environment, Akamatsu (2017) presents a "vocalization strategy" in which words are memorized by listening to the sound of the word and also saying it out aloud. Using path analysis, Akamatsu determined that the strategy correlates academic performance in English. In this regard, Krashen's (1987) input theory states that in SLL, repeated input improves learners' language proficiency. Concurrently, Swain (1998) pointed out that outputting knowledge improves learners' linguistic formality by raising their awareness of language formality. Vocalization strategy is considered to be effective in improving phonetic knowledge of English vocabulary because it is an approach to learn words while simultaneously inputting and outputting the sounds of words through audio equipment. Akamatsu indicates a utilization-oriented view of learning that emphasizes daily use and conversation in English as a factor that defines the use of vocalization strategy. However, practical research on supporting the acquisition of learning strategies has indicated that intervention in learning perspectives and the teaching of target strategy knowledge alone are not effective for strategy retention (e.g., Ueki 2004). Jenkins (2000) also argues that efforts to improve speech habits will not be made unless learners understand the need to do so.

Correspondingly, a mechanism is required for learners to emphasize the importance of phonetic information of words when learning vocabulary, and to be aware of its improvement. Simultaneously, a mechanism to promote the use of vocalization strategy is also required.

1.2 Metacognition

One of the factors that can be utilized in promoting the use of vocalization strategy is metacognition. It is defined as the awareness or knowledge that allows learners to understand their learning strengths and weaknesses, the cognitive resources available to meet the demands of the task, and to adjust their learning behavior to optimize the learning process and outcomes (Winne and Perry 2000). Through metacognition, learners set achievement goals in learning tasks, select appropriate learning strategies to achieve them, and monitor, evaluate, and adjust their learning behaviors (Haslaman 2011). Many studies have shown that enhancing metacognition promotes the use of learning strategies (e.g., Akin et al. 2007).

As a mechanism for enhancing metacognition in SLL, Swain (1998) suggests that "output activity" is a factor that enhances learners' awareness of a language form. Furthermore, it has been suggested that dictation-type speech recognition, which recognizes learners' speech and transcribes it into letters, may promote metacognitive strategies such as awareness and self-monitoring (Dai and Wu 2021). Since such speech recognition technologies do not accept inappropriate pronunciations, it is expected to be effective in making learners aware of the importance of speech sounds, activating metacognition, and encouraging the use of vocalization strategy. Additionally, metacognitive support (MS) using Learning Analytics (LA) can also play a crucial role in metacognitive activation (Yilmaz and Yilmaz 2020). LA is defined as the measurement, collection, and analysis of learner data to understand and optimize learning, and based on the data analyzed in LA, learners can monitor, improve, and evaluate their learning behaviors and learning outcomes (Durall and Gros 2014; Yilmaz and Yilmaz 2020). MS is an educational approach that supports the improvement of metacognition (e.g., Jumaat and Tasir 2016; Ku¨nsting et al. 2013; Molenaar, I. et al. 2014; Schwonke et al 2013; Yilmaz and Keser 2017). Essentially, questions and inquiries that promote awareness and improvement of learning behaviors have been proposed as methods of MS. Yilmaz and Yilmaz (2020) have demonstrated that LA can activate learners' metacognition by creating an environment in which they can monitor their learning data, while MS can encourage them to plan, monitor, and evaluate their learning behaviors.

Accordingly, this study will develop a web-based English vocabulary learning system that utilizes speech recognition technology and LA. Through the system, learners learn English words and pronounce them. The pronunciation is then transcribed using dictation-type speech recognition. By presenting the transcribed data and LA data of the learning behavior, the system improves learners' metacognition of phonetic knowledge and encourages the use of vocalization strategy in English vocabulary learning. Ultimately, the goal is to improve learners' vocabulary in the phonetic aspect.

2. BACKGROUND

2.1 Speech Recognition

In pronunciation learning, feedback on learners' pronunciation is considered important. The use of automatic speech recognition has been attracting attention as a means of providing this feedback. Automatic speech recognition recognizes learners' speech on a computer and provides feedback in a timely manner based on the correctness or incorrectness of the speech (Neri et al. 2008). Dictation-type speech recognition, which does not score speech or identify errors, has been suggested to promote metacognitive strategies such as awareness and self-monitoring in learners (Dai and Wu 2021). Pronunciation learning using dictation-type speech recognition requires learners to identify pronunciation errors based on dictation texts (Liakin et al. 2015). In the process, learners may reflect on the problems of speech form and learning methods. Previous studies have indicated that dictation-type speech recognition contributes to the improvement of learners' autonomy in learning activities as well as pronunciation skills (Liakin et al. 2015; Mroz 2018; Dai and Wu 2021). By enhancing learner autonomy, learning activities such as listening and pronunciation of sounds of words in vocabulary learning is practiced, leading to the improvement of not only speech but also pronunciation recognition, which is believed to improve vocabulary comprehensively in the phonetic aspect. Unfortunately, prior research on automatic speech recognition has focused only on its relationship with pronunciation ability. These studies have not clarified the relationship between automatic speech recognition and learning strategies. While automatic speech recognition has been shown to be effective in improving speech knowledge, it is necessary to examine the impact of this technology on learning methods and behaviors. By building a speech recognition system with a learning page and recording vocabulary learning behavior, it is possible to clarify the relationship between speech recognition and learning behavior, which has not been elucidated in previous studies.

2.2 Metacognitive Support by LA

Metacognitive support (MS) helps learners improve their self-regulated learning skills by offering an objective view of their learning behaviors and learning outcomes (e.g., Jumaat and Tasir 2016; Ku¨nsting et al. 2013; Molenaar et al. 2014; Schwonke et al. 2013; Yilmaz and Keser 2017). In online learning environments, Learning Analytics (LA) is an effective tool for metacognitive support. Chen et al (2019) suggested an LA design based on metacognition, and Chen et al (2020) indicated that visualization of learning behaviors improves certain aspects of metacognition. Yilmaz (2014) reported that providing MS online improves students' metacognition. In addition, LA enables learners to understand their learning process and review their learning behavior by visualizing the results of the analysis of their learning behavior as data. Therefore, LA can activate metacognition and promote the use of effective learning strategies by enabling learners to monitor their own learning behaviors and learning outcomes. Based on this theory, it is expected that presenting learners with dictation texts recognized by speech recognition and LA learning data in vocabulary learning could help them identify problems in their own phonetic knowledge and autonomously use a phonetic memory strategy that emphasizes the sound of words as a method to learn words. This is expected to encourage learners to identify problems with phonetic knowledge and voluntarily use the vocalization strategy that emphasizes the sound of the word as a word-learning method.

3. AUTOMATIC SPEECH RECOGNITION

This study uses Mozilla Developer Network's (MDN) Web speech API as the automatic speech recognition method. Web speech API recognizes user speech and transcribes it into text. There are many speech recognition tools other than Web speech API, such as Apple's Siri, Amazon's Alexa, Microsoft's Cortana, and Google Assistant. Mroz (2018) presents four criteria for the selection of automatic speech recognition: cost, accessibility, familiarity, and technology. Web speech API can be implemented for free by using JavaScript, and can also be used with Chrome, which has the highest market share in Japan among web browsers (StatCounter 2021–2022). With regard to accuracy, Ashwell and Elam (2017) investigated the accuracy rate of

Google Web Speech API for utterances based on the speech of native English speakers and Japanese non-native speakers. According to the study, the recognition rate of Google Web Speech API is 89.4% for native English speakers and 65.7% for non-native speakers. While the study points out that the recognition rate declines for proper nouns and certain collocations, the high recognition rate for native speakers' speech highlights the advantages of using Web Speech API for language learning to encourage reflection on speech. In addition, since Web Speech API does not correct pronunciation errors according to the user's native language, it is possible for learners to identify pronunciation errors by referring to the recognition results (Evers and Chen 2020). Therefore, the use of Google Web Speech API as a speech recognition tool in this study was found to encourage Japanese learners' reflection on their phonetic knowledge of English words.

4. THE DESIGN OF WEB-BASED APPLICATION

4.1 Outline of the Application

The system consists of a learning process (see Figure 1), an output process (see Figure 2), and a self-assessment process (see Figure 3) for English vocabulary learning. First, learners enter a learning process and engage in English vocabulary learning by comparing the English word with its translation. After learning vocabulary, learners move to the output process to pronounce the English words learned on the learning page. In the output process, the speech recognition system displays the dictation text, which is then checked against the word information to determine if they are correct or incorrect. If all the words are pronounced correctly, the learner is redirected to the learning page of the next lesson. However, even if one word is uttered incorrectly, the learner is redirected to the self-assessment process to reflect his/her learning outcome on the output page and learning strategy on the learning page based on MS by LA. When the learner has answered all the questions correctly, the system moves to the next lesson. This cycle is repeated to encourage learners to understand the importance of audio information of vocabulary and encourage the use of vocalization strategy on the learning page.

Figure 1. The user interface of learning process

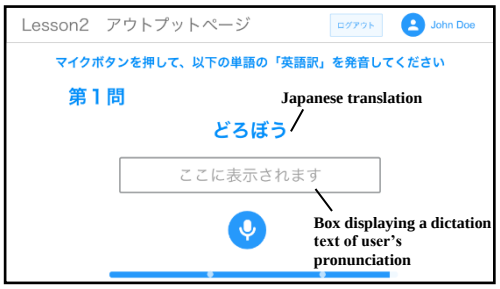

Figure 2. The user interface of output

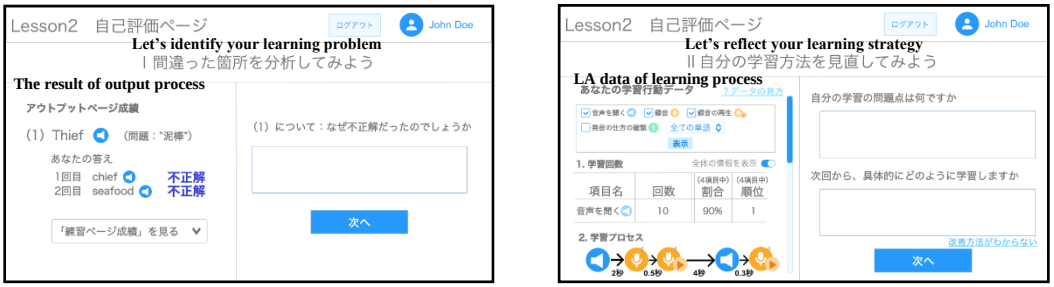

Figure 3. The user interface of self-assessment process processs

4.2 Learning Process

The learning page displays the English word and its Japanese translation, and the user learns English vocabulary by comparing English words along with the translation. In addition, the learning page includes an "audio function" to verify the sound of each word, a "recording function" to record and play back one's own pronunciation, and a "pronunciation point confirmation function" to explain how to pronounce the word. The audio function uses JavaScript's voice reading function, and by pressing a button, the user can listen to the voice of the word. The playback speed can be adjusted by the user from 0.8x to 2x (Dai and Wu 2021). The "recording function" will be implemented in Vue.js. The recording can be started and stopped by a push button. The recorded audio is displayed on the page as an audio file, and users can play back the recorded audio as long as they like. The audio file can be deleted and re-recorded. Clicking the "Next" button on the study page takes the user to the "Output" page.

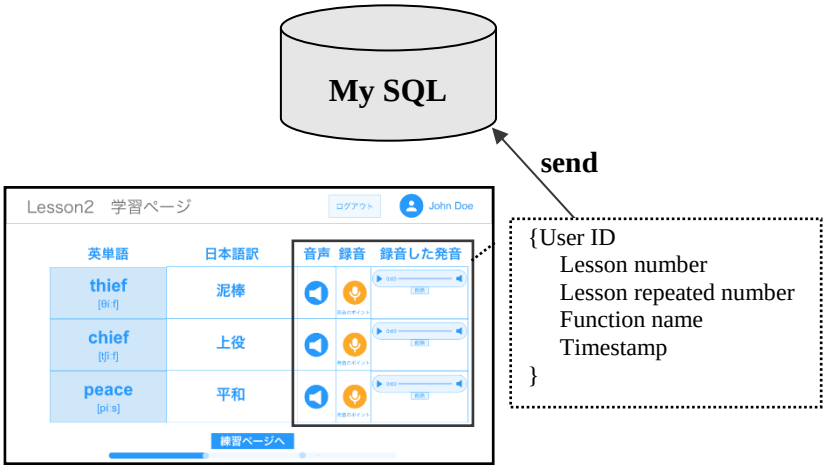

Figure 4. The learning log accumulation process

4.3 Output Process

On the output page, the user pronounces the English words learned from the learning page. The speech recognition system recognizes the words uttered and displays the dictation text (see Figure 5). The recognition results are verified against the word information to determine if they are correct or incorrect (see Figure 6). Web Speech API, a free JavaScript library, is used as the speech recognizer. The user can pronounce each word up to two times; if the first pronunciation is correct, the user moves on to the next question; if both the words are pronounced incorrectly, the user cannot retry the same question and moves on to the next question. After all the words have been pronounced, the results page displays the pass/fail results for all words: "Pass" if the first or second pronunciation was appropriate, and "Fail" if the first or second pronunciation was inappropriate. If the learner answers all the questions correctly, the user is redirected to the study page of the next lesson; even if one question is incorrect, the user is redirected to the "Self-Evaluation Page."

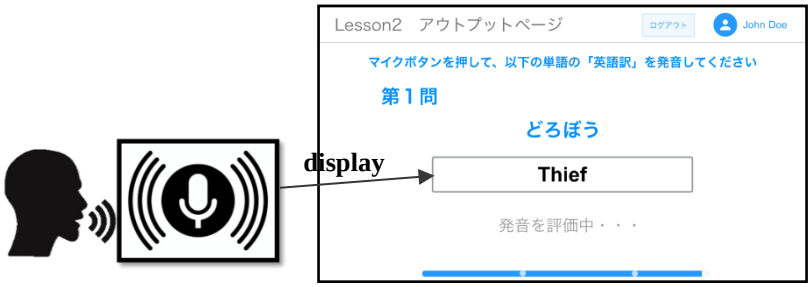

Figure 5. Displaying dictation text process

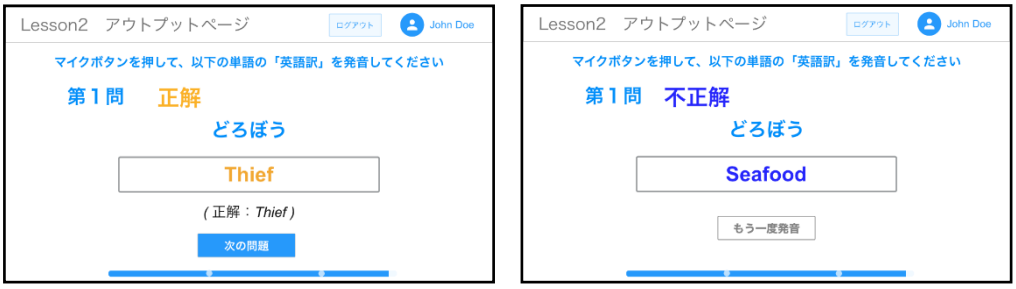

Figure 6. Right and wrong judgment

4.4 Self-Assessment Process

On the self-assessment page, the user reflects the results on the output page and the learning strategy for vocabulary learning by answering questions based on MS (Yilmaz and Keser 2017). At that time, the learning log on the learning page and the results log on the output page are displayed as LA data. To promote learners' metacognition, the self-assessment process encourages the user to identify learning problems and review the learning strategy (Akin et al. 2007; Schraw and Dennison 1994; Schraw and Moshman 1995). First, to promote an understanding of learning problems, questions judged as "Fail" on the output page are extracted, and texts and audio data of the correct answers and the recognition results are presented as a reflection of individual questions (see Figure 7). To improve metacognition, the question, "Why do you think you made a mistake?" is presented as MS, and once the answer is entered in the text box, the user is reassigned to the next section. This section presents the number of times the user has used the audio or recording functions on the learning page, which helps review his/her learning strategy based on this data (see Figure 8). At that time, the questions, "What learning method should I review?" and "How should I improve my learning method?" are presented as MS, and once the responses are entered in the text boxes, the user is transferred to the self-assessment confirmation page. On the confirmation screen, the results of self-assessment are displayed as a list, and after checking them, the user is redirected to the learning process of the same content. If a student answers correctly a question that was marked as failed on the previous output page, the user is prompted to consider the reasons why he/she was able to answer correctly and the effective learning methods on the learning process. Specifically, the user is presented with the questions, "Why do you think you were able to answer correctly?" and "What learning methods do you think were effective? When the learner has answered all the questions correctly, the system moves to the learning page of the next lesson.

Figure 7. The process of identifying learning problems

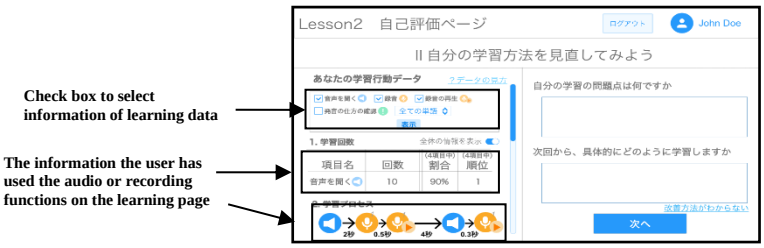

Figure 8. The process of reviewing learning strategy

5. CONCLUSION

This paper describes the need for an English vocabulary learning system that supports the acquisition of vocalization strategy and its system design. After developing the system and formatively evaluating its validity, we would like to examine whether this system will promote the acquisition of vocalization strategy and contribute to the improvement of learners' vocabulary.

ACKNOWLEDGEMENT

This work was supported by JSPS Grants-in-Aid for Scientific Research JP22H00552 and JP21K18134.

REFERENCES

Akamatsu, D., 2017. Relation Between High School Students' Beliefs and Learning Strategies, and Their Academic Achievement in Learning English. *Japanese Journal of Educational Psychology*, Vol. 65, pp. 265–280

Akin, A. et al, 2007. The validity and reliability of the Turkish version of the metacognitive awareness inventory. *Educational Sciences: Theory and Practice*, Vol. 7, No. 2, pp. 671–678

Ashwell, T. and Elam, J.R., 2017. How accurately can the Google Web Speech API recognize and transcribe Japanese L2 English learners' oral production? *The JALT Call Journal*, Vol. 13, No. 1, pp. 59–76

Chen, L. et al, 2019. Design of learning analytics dashboard supporting metacognition. *Proceedings of 16th International Conference Cognition and Exploratory Learning in Digital Age (CELDA 2019)*. Cagliari, Italy, pp. 175–182.

Chen, L. et al, 2020. Factors of the use of learning analytics dashboard that affect metacognition. *Proceedings of 17th International Conference Cognition and Exploratory Learning in Digital Age (CELDA 2020)*. Lisbon, Portugal, pp. 295–302

Dai, Y. and Wu, Z., 2021. Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: a mixed-methods study. *Computer Assisted Language Learning.*

Durall, E. and Gros, B., 2014. Learning analytics as a metacognitive tool. *Proceedings of 6th International Conference on Computer Supported Education (CSEDU).* Barcelona, Spain, pp. 380–384.

Evers, K. and Chen, S., 2020. Effects of an automatic speech recognition system with peer feedback on pronunciation instruction for adults. *Computer Assisted Language Learning*, pp. 1–21.

Fan, M. Y., 2003. Frequency of use, perceived usefulness, and actual usefulness of second language vocabulary strategies: A study of Hong Kong learners. *The Modern Language Journal*, Vol. 87, No.2, pp. 222–241.

Field, J., 2004. An insight into listeners' problems: Too much bottom-up or too much top-down? *System*, Vol. 32, pp 363–377.

Haslaman, T., 2011. *Effect of an online learning environment on teachers' and students' self-regulated learning skills* (Doctoral dissertation). Hacettepe University, Ankara, Turkey.

Huo, S. and Wang, S., 2017. The effectiveness of phonological-based instruction in English as a Foreign Language students at primary school level: A research synthesis. *Frontiers in Education*, Vol. 2, pp. 1–13.

Jenkins, J., 2000. *The Phonology of English as an International Language*. Oxford University Press, Oxford, UK.

Jumaat, N. F. and Tasir, Z., 2016. A framework of metacognitive scaffolding in learning authoring system through Facebook. *Journal of Educational Computing Research*. Vol. 54, No. 5, pp. 619–659.

Krashen, S. D., 1987. *Principles and Practice in Second Language Acquisition*. Pergamon Press, Oxford, UK.

Ku¨nsting, J. et al., 2013. Enhancing scientific discovery learning through metacognitive support. *Contemporary Educational Psychology*, Vol. 38, No. 4, pp. 349–360.

Li, S. and Woore, R., 2021. The effects of phonics instruction on L2 phonological decoding and vocabulary learning: An experimental study of Chinese EFL learners. *System*, Vol.103

Liakin, D. et al., 2015. Learning L2 pronunciation with a mobile speech recognizer: French/y/. *CALICO Journal*, Vol. 32, No. 1, pp. 1–25.

Molenaar, I. et al., 2014. Metacognitive scaffolding during collaborative learning: A promising combination. *Metacognition and Learning*, Vol. 9, No. 3, pp. 309–332.

Mroz, A., 2018. Seeing how people hear you: French learners experiencing intelligibility through automatic speech recognition. *Foreign Language Annals*, Vol. 51, No. 3, pp. 617–637.

Nation, I. S. P., 2001. *Learning Vocabulary in Another Language*. Cambridge University Press, Cambridge, UK.

Neri, A. et al., 2008. The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning*, Vol. 21, No. 5, pp. 393–408.

Odlin, T., 1989. *Language Transfer: Cross-linguistic Influence in Language Learning*. Cambridge University Press, Cambridge, UK.

Ota, M. et al, 2009. The KEY to the ROCK: Near-homophony in nonnative visual word recognition. *Cognition*, Vol. 111, pp. 263–269.

Oxford, R. L., 1990. *Language Learning Strategies: What Every Teacher Should Know*. Newbury House, New York, USA.

Rasouli, F. and Jafari, K., 2016. A Deeper Understanding of L2 Vocabulary Learning and Teaching: A Review Study. *International Journal of Language and Linguistics.* Vol. 4, No. 1, pp. 40–46.

Schmitt, N., 1997. Vocabulary learning strategies. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition, and Pedagogy*. Cambridge University Press, Cambridge, UK, pp. 199–227.

Schwonke, R. et al, 2013. Metacognitive support promotes an effective use of instructional resources in intelligent tutoring. *Learning and Instruction*, Vol. 23, pp. 136–150.

Schraw, G. and Dennison, R. S., 1994. Assessing metacognitive awareness. *Contemporary Educational Psychology*, Vol. 19, No. 4, pp. 460–475.

Schraw, G. and Moshman, D., 1995. Metacognitive theories. *Educational Psychology Review*, Vol. 7, No. 4, pp. 351–371.

Shimamoto, T., 2002. Why it is necessary to increase vocabulary: the relationship between vocabulary and English language skills. *The English Teachers' Magazine*, Vol. 50, No. 12, pp. 8–10.

Statcounter Global Stats, 2021–2022. *Browser Market Share Japan Sept 2021 – Sept 2022*,

https://gs.statcounter.com/browser-market-share/all/japan

Swain, M., 1998. Focus on form through conscious reflection. In C. Doughty & J. Williams (Eds.), *Focus on Form in Classroom Second Language Acquisition.* Cambridge University Press, New York, USA, pp. 64–81.

Teshima, M., 2011. On Teaching English Pronunciation in Japanese Secondary Education How It Is and How It Should Be. *Journal of the Phonetic Society of Japan*, Vol. 15, No. 1, pp 31–43.

Ueki, R., 2004. Ideal Ways to Teach Students How to Utilize Self-Monitoring Strategies: Beliefs About Learning and Knowledge About Strategies. *Japanese Journal of Educational Psychology*, Vol. 52, pp. 277–286

Winne, P. H. and Perry, N. E., 2000. Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich & M. Zeidner (Eds.), *Handbook of Self-regulation* (Chapter 16). Academic Press, San Diego, USA..

Yilmaz, F. and Yilmaz, R., 2020. Learning analytics as a metacognitive tool to influence learner transactional distance and motivation in online learning environments, *Innovations in Education and Teaching International*.

Yilmaz, R. and Keser, H., 2017. The impact of interactive environment and metacognitive support on academic achievement and transactional distance in online learning. *Journal of Educational Computing Research*, Vol. 55, No. 1, pp. 95–122.

Yilmaz, R., 2014. *The effect of interaction environment and metacognitive guidance in online learning on academic success, metacognitive awareness and transactional distance* (Doctoral dissertation). Ankara University, Ankara, Turkey.