**English Vocabulary Learning Software Developed by Students at a National Institute of Technology (KOSEN)**

M.K. Higa\*,a, Y. Ashidab , T, Kawakamic, Y. Koyamad, S. Teshimab, and S. Yamane b

a Institute for Foreign Language Research and Education, Hiroshima University, Hiroshima, Japan

b Department of Electrical Engineering and Computer Science, National Institute of Technology, Matsue College, Matsue, Japan

c Graduate School of Advanced Science and Engineering, Hiroshima University, Hiroshima, Japan

d Department of Electrical Engineering and Computer Science, National Institute of Technology, Matsue College, Matsue, Japan

\*mhiga@hiroshima-u.ac.jp

**Abstract**

**This article reviews a software development project aimed at assisting EFL learners in analyzing texts and creating personalized word lists. These word lists are stored in a cloud system for later review. The article begins by discussing the pedagogical rationale behind the system, referring to relevant literature and concepts related to second-language vocabulary acquisition. Following that, a demonstration of the software's usage is presented, highlighting its key features. The final section of the article explores various technical aspects of the platform and how they contribute to enhancing vocabulary acquisition. This collaborative project involves instructors and students from a National Institute of Technology (KOSEN), who have contributed their programming expertise and valuable feedback as learners to optimize the software for their own studying needs.**

**Python is the primary programming language for this project, and uses Python libraries, Django and the Natural Language Toolkit (NLTK). These libraries enable users to perform text analyses and gather information such as word frequency, difficulty, and genre. To provide learners with more specific details about encountered words, the text-analysis tool incorporates data from commonly used word lists as well as custom lists developed within the researchers' institutions. To facilitate efficient review, students can save selected vocabulary in a spaced repetition system (SRS). SRS is a learning tool that presents information at increasing intervals based on the user's performance, ensuring effective long-term retention. It is based on the Leitner System, which organizes flashcards into boxes reviewed at progressively wider intervals to maximize retention. SRS utilizes algorithms to determine the optimal review interval for vocabulary words, promoting efficient retention.**

**Although similar web-based tools exist, the combination of text analysis functions and SRS makes this software unique, providing learners with a flexible and convenient way to personalize their vocabulary lists. This feature will be particularly beneficial for learners focusing on specialized fields such as engineering. By utilizing domain-specific word lists and vocabulary, students can better prepare themselves to use English in their future professional endeavors.**

**Keywords:** *second language acquisition, data-driven learning, spaced-repetition software, autonomous learning, English for specific purposes*

**Introduction**

 Vocabulary acquisition is foundational in language learning. Wilkins (1972) famously wrote, “While without grammar very little can be conveyed, without vocabulary nothing can be conveyed (p. 111). In regions where incidental vocabulary acquisition is typically unlikely due to limited opportunity to interact in the target language, an intentional vocabulary learning method is very important. Advances in technology have greatly aided this process and in the present, language learners have access to more computer-based vocabulary study tools than one could ever possibly utilize in a lifetime.

 The creation of the software presented in this paper was achieved through the collaborative efforts of instructors and students at a National Institute of Technology (KOSEN). It considers various challenges that students may face in language learning, and draws upon on their skills in computer programming to find solutions to these challenges.

 A common challenged that Japanese EFL learners face, is the lack of adequate time to review materials as they matriculate up through the education system, leading to gaps in their proficiency. With pressures stemming from entrance exams and standardized tests, there can be a disconnect between the goals and reality. MacWhinnie and Mitchell (2021) contend that the extensive amount of grammar knowledge necessary to score well on university entrance exams leaves limited opportunity to review and engage with materials previously studied. Furthermore, assuming that even when a curriculum is well-connected between various levels, there is a need for students to be able to manage materials previously studied, ideally into one, well-maintained database that is customizable for their aims. The software being developed aims to integrate reading materials, vocabulary lists and corpus software into one, easy-to-use platform that will help Japanese EFL learners improve their learning of vocabulary words.

 There are a vast number of computer-assisted language learning mediums to achieve this aim. Currently however, vocabulary acquisition software, extensive reading resources and corpus software exist mainly as separate entities. In Japan, one of the favored methods of vocabulary study is through rote memorization of word lists (Yamamoto, 2014). Reading skills (and to an extent, vocabulary is included in this) are developed in various ways, such as through the use of graded readers, or textbooks, and in recent years, utilizing corpus software. Though students now have access to software made for all three of these areas of study, they could benefit to an even greater extent if these programs were combined into one, integrated platform. This project details the creation of new software that combines aspects of currently available software to provide a practical and efficient learning tool. We will explain how the proposed system has been conceived through research-based reasoning.

**Pedagogical Background**

Selection of word lists

 The selection of vocabulary items for personalized word lists requires careful consideration and informed decision-making based on the learner's needs. For instance, research suggests that focusing on the most frequently used 100 words in English can cover up to 50% of the words found in texts (Nation, 2016). While this information may be well-known to academics, it might be unfamiliar to Japanese EFL learners in tertiary education and below. Therefore, making such information easily accessible to students can help them make better choices for self-study. Word lists, both general and specific, play a crucial role in providing learners with valuable information for selecting words to study. For example, the New General Service List consists of 2,800 vocabulary items, covering 92% of most English-based texts (Browne et al., 2013). Other lists, such as the CEFR-J word list, cater to the specific needs of Japanese EFL textbooks, and the New JACET 8000 categorizes words by difficulty level (Ishikawa, n.d.). These lists, along with others, are included in the software being developed by the authors. A description of the word lists being utilized are described below:

TABLE 1. Selected Wordlists Utilized in the Program (Higa & Ashida, 2023)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Word list | Type of coverage | Reason for selection | Number of words | Source |
| NGSL 1.1 | Covers high frequency English words | Covers 92% of the most frequent words in English texts  | 2800 | Browne, C., Culligan, B. & Phillips, J. (2013) |
| CEFR-J | Corpora of English textbooks used in Japan and neighboring countries  | Applies CEFR standards to categorize words by difficulty | 7801 | Tono, 2022 |
| TOEIC Service | Provides 99% of TOEIC test coverage when combined with NGSL | TOEIC is the most taken test in Japan | 1200 | Browne, C. and Culligan, B. (2016) |
| New JACET 8000 | Educational word list for Japanese University students | Specifically designed for Japanese EFL learners | 8000 | Ishikawa, N.D. |
| Hirotan | Covers basic and advanced words for daily conversation, academics, and business | Word list used in general English classes at Hiroshima University and contains original example sentences | 6000 | Enokida et al., 2018 |
| MEWL | A word list for medical purposes, it focuses on body systems | Word list used in a medical course at Hiroshima University | 1750 | Davies et al., 2020 |

Text-analysis for EFL learners

 An additional feature we included in the software was a text-analysis tool to allow learners to perform basic analyses of words they encounter in texts. In our software, this feature allows for students to determine whether or not a word is worth reviewing based on basic data, such as the word lists they are included in, and the difficulty of the word. This is in contrast to the more common method of word memorization in Japan, which is to study word lists aimed at students taking standardized tests. Thomson and Mehring (2016) provided a comprehensive overview of numerous studies supporting the prevalence of rote memorization as the primary vocabulary learning approach among Japanese students. Although this method may prove effective in achieving short-term goals such as passing tests, the lack of contextualization and interaction with words in natural settings makes it difficult to retain these vocabulary items over the long-term. By leveraging simple text-analysis tools, students may enhance their ability to comprehend words at a deeper level, thereby promoting more effective language learning.

Autonomous learning

 By equipping learners with the necessary understanding to make well-informed choices regarding word list compilation and basic text analysis, learners will be empowered to engage in self-directed learning beyond formal educational environments. Autonomous learning entails the learner’s ability to self-monitor their own learning process and make decisions regarding what they need to learn, how to learn it, and how to apply it. According to Schwienhorst (2007), learner autonomy is characterized as “a learner-centered approach to learning, encouraging learners to critically reflect on their learning process and establish a personally meaningful connection with it” (p. 11). This approach proves particularly advantageous in second language acquisition as it allows learners to customize their learning experience according to their unique needs, interests, and relevant areas of focus. Furthermore, it addresses the issue of maintaining continuity throughout different stages of study.

 Retaining knowledge and educational materials from an institution for self-study purposes after leaving formal education presents a challenge in language learning. This challenge becomes especially pertinent when English proficiency is required outside of an academic setting. The research and development of engineering English materials conducted by the first author have highlighted the absence of an organized methodology among professionals to apply the knowledge acquired during their tertiary education to their respective fields. One objective of this software is to bridge this gap by enabling users to store and access study materials and data for future review and study.

Data-driven learning (DDL)

 Data-driven learning (DDL) is an approach that utilizes data to inform and guide the learning process, empowering individuals to gather data and make informed decisions about their own study path. Szudarski (2017) emphasizes that DDL not only raises learners’ awareness of real-life language usage but also fosters autonomy by encouraging them to take responsibility for their learning process. Furthermore, our software’s DDL approach facilitates a shift from passive to active learning as it encourages learners to actively engage with texts and think critically while reading.

 Additionally, while general word lists are essential for foundational learning, some students may have specific goals, such as their future profession or a particular academic context, for learning English. Access to text analysis tools allows learners to make informed choices about what to study based on their individual needs. Anthony (2018) suggests that while general word lists provide valuable information, learners can acquire more specific knowledge by conducting their own text analysis. Browne et al. (2013) support this idea, advising learners that studying vocabulary beyond the NGSL 1.1 list is best achieved by focusing on words within their specialization. They state, “The number of words they need to learn to make an additional 1% coverage gain increases sharply after 92%, and… depending on the student’s specialization, it is very likely that they will make significantly faster gains by learning special purpose vocabulary” (n.d.). This viewpoint highlights the importance of tailoring vocabulary studies to specific needs and interests. Our aim is to facilitate this process by enabling users to compile personalized word lists.

**Software Development**

 The software developed in this project has been provisionally called the “Hi-lex System” (short for *Hiroshima-lexicon system*). It is composed of two primary elements: a user interface (frontend), which encompasses the visible aspects on the website, and a server-side component (backend), responsible for handling user accounts, storing word lists, and generating data accessible to researchers and site administrators. Here’s an overview of the functions users have access to:

TABLE 2. Main Functions of software

|  |
| --- |
| *Users of this software will be able to:* |
| 1. choose texts (autonomous learning) |
| 2. perform vocabulary profiling analysis based on selected word lists (Data-driven learning) |
| 3. compile and save personalized word lists (autonomous learning) |
| 4. review words efficiently (Spaced repetition) |
| 5. store reading materials for future review |

A visual for the process of how the software is used can be found below:



FIGURE 1. Visualization of the Vocabulary Selection Process (Higa & Ashida, 2023)

 The software takes advantage of the open-source data analysis library in Python, called the Natural Language Toolkit (NLTK), designed to process human language data. It was created in the late 1990s by Steven Bird and Edward Loper at the University of Pennsylvania to aid in natural language processing (NLP)and is able to perform lexical tasks such as tokenization, part-of-speech tagging, parsing, semantic reasoning, and sentiment analysis. NLTK provides developers with to access to vast linguistic datasets. It is distributed under the Apache 2.0 open-source license, which allows users to freely use, modify, and distribute the NLTK codebase.

 The front end of the application was developed with another Python library, Django, a Python web framework that provides a set of libraries for web development. Django includes features such as form handling, user authentication, and an administration interface, which are utilized in our application. With Django, developers can build dynamic and scalable web applications quickly, thanks to its comprehensive set of built-in functionalities.

 The design was achieved by using Bootstrap, a front-end framework that provides a collection of pre-designed CSS and JavaScript components. It aids in the development of responsive web interfaces by offering a grid system, typography, navigation bars, buttons, forms, and other user interface (UI) elements that can be integrated into web projects.

 In order to review vocabulary items chosen by the student, a spaced-repetition system (SRS) is used. SRS is a learning technique that optimizes review schedules by spacing out the intervals between learning sessions based on the difficulty and retention of the material, enhancing long-term memory retention. The SRS component of the software utilizes the SuperMemo2 algorithm, developed by Piotr Wozniak, a Polish computer scientist and founder of the SuperMemo World company. SuperMemo2 calculates an optimal review interval for each item based on the user’s historical performance in recalling that item. When an item is correctly recalled, the interval until the next review is extended, allowing for longer gaps between reviews. Conversely, if an item is forgotten, the interval is shortened, leading to more frequent reviews. This adaptive algorithm aims to strike a balance between challenging the learner with timely reviews and avoiding overwhelming them with excessive repetitions. By systematically adjusting the review intervals, SuperMemo2 maximizes memory retention and helps learners efficiently retain information over the long term.

At the present, the front end of the software can be seen in the screenshot below:

****

FIGURE 2. Screenshot of Hi-lex System

**Discussion**

Talk about using SuperMemo 2 and how it could be edited later on depending on feedback about the algorithm. At the time of this writing, the authors have not yet conducted a study on the effectiveness of the software, nor have they received formal feedback from users. This will

At the time of this writing, the software is hosted locally, and not available on the internet. However, upon completion of the software, it will be deployed on a cloud system to enable users to use the software on multiple devices.

**Conclusion**

 This paper has surveyed some of the innovations in language learning tools related to second language vocabulary acquisition and selected various elements of this software that can serve the constituents the authors are most familiar with, in this case, Japanese students studying engineering. We have reviewed some of the current online corpus and vocabulary acquisition tools available for self-study and hope to add to this current body of software based on the specific needs of the students we are familiar with. As practitioners and learners, we hope that the software being developed in this project will enable students to take more agency in the selection and compilation of their word lists, engage with texts more actively, and make their studying more efficient.

**Acknowledgements**

 Each author of this paper has had the privilege of being a part of Matsue Kosen, either as students or faculty members. We want to thank everyone at Matsue Kosen for creating the special environment that enabled us to develop this program. This work was supported by JSPS KAKENHI Grant Number JP20K13155.

**References**

Anthony, L. (2017). *AntConc* (Version 3.5.0) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from www.laurenceanthony.net/software/antconc/

Anthony, L. (2022). *AntWordProfiler* (Version 2.0.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Anthony, L. (2018). *Introducing English for Specific Purposes.* Routledge.

Browne, C., Culligan, B., & Phillips, J. (2013). The New General Service List. Retrieved from http://www.newgeneralservicelist.org.

Browne, C., & Culligan, B. (2016). The TOEIC Service List. Retrieved from http://www.newgeneralservicelist.org.

Brown, D., & Stoeckel, T. (2019). *VocabLevelChecker*. VocabLevelChecker. Retrieved December 23, 2022, from https://vocablevelchecker.net/index.html

Cheng, W. (2012). *Exploring Corps Linguistics: Language in action.* Routledge*.* https://doi.org/10.1111/j.1540-4781.2013.12052\_13.x

Chun, B. A., & Heo, H. J. (2018). The effect of flipped learning on academic performance as an innovative method for overcoming Ebbinghaus’ forgetting curve. *Proceedings of the 6th International Conference on Information and Education Technology* (pp. 56–60).

Cobb, T. Compleat Lexical Tutor [computer program]. Accessed 18 Dec 2022 at https://www.lextutor.ca . Cobb, T., (2004). The Compleat lexical tutor, v. 4. *TESL-EJ, 8*(3).

Davies, W., Fraser, S., Tatsukawa, K., & Enokida, K. (2020). Reflections on an EMP (English for Medical Purposes) project. *Hiroshima Studies in Language and Language Education*, 201–216.

Duolingo. (2022). *Duolingo* [Web application]. Retrieved December 23, 2022, from https://www.duolingo.com/

Enokida, K., Kusanagi, K., Kida, S., Morita, M., & Sakaue, T. (2018). Tracking online learning behaviour in a cross-platform web application for vocabulary learning courses. *Future-proof CALL: Language Learn- ing as Exploration and Encounters*, 47.

Gu, P. Y. (2003). Vocabulary learning in a second language: Person, task, context and strategies. *TESL-EJ, 7*(2), 1–25.

Hamada, A., Iso, T., Kojima, M., Aizawa, K., Hoshino, Y., Sato, K., ... & Yamauchi, Y. (2021). Development of a vocabulary size test for Japanese EFL learners using the New JACET List of 8,000 basic words. *JACET Journal*, *65*, 23–45.

Higa, M., & Ashida, Y. (2023). A Pedagogical Framework for the Development of Software for EFL Learners. *Hiroshima Studies in Language and Language Education*, 63–77.

Ishikawa, S. (n.d.). 神戸大学石川慎一郎研究室/*Dr. Shin Ishikawa, Kobe University*. Lab of Dr. Shin Ishikawa, Kobe University. https://language.sakura.ne.jp/s/voc.html

Key Points of the Revised Foreign Language National Curriculum Standards (Course of Study) at Elementary School and Lower Secondary School Levels. (2021, June 23). *National Institute for educational policy research*. Retrieved December 18, 2022, from https://www.nier.go.jp/English/educationjapan/

MacWhinnie, S.G.B., & Mitchell, C. (2017) English classroom reforms in Japan: a study of Japanese univer- sity EFL student anxiety and motivation. *Asian-Pacific Journal of Second and Foreign Language Edu- cation, 2:7,* 1–13. https://doi.org/10. 1186/s40862-017-0030-2

Memrise. (2022). *Memrise* [Web application]. Retrieved December 23, 2022, from https://memrise.com

Meyer, C. (2002). *English Corpus Linguistics: An Introduction* (Studies in English Language). Cambridge

University Press. doi:10.1017/CBO9780511606311

Microsoft. (n.d.). *What is cloud computing? A beginner’s guide*. Cloud Computing Services | Microsoft Azure. Retrieved December 23, 2022, from https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-cloud-computing/

Mizumoto, A. (2021). New Word Level Checker [Web application]. https://nwlc.pythonanywhere.com/

Nation, P. (2013). *Learning Vocabulary in Another Language* (2nd ed., Cambridge Applied Linguistics). Cambridge University Press. doi:10.1017/CBO9781139858656

Nation, P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company. doi: https://doi.org/10.1075/z.208

Okamoto, M. (2007). Lexical attrition in Japanese university students: A case study. *JACET Journal*, (44),

*Purposes of the CEFR*. (n.d.). Common European Framework of Reference for Languages (CEFR). https://www.coe.int/en/web/common-european-framework-reference-languages/uses-and-objectives

Römer, U. (2010). Using general and specialized corpora in English language teaching: Past, present, and future. In M. C. Campoy-Cubillo, B. Bellés-Fortuño, & L. Gea-Valor (Eds.), *Corpus-based approaches to English language teaching (pp. 18–35).* Continuum.

Schwienhorst, K. (2007). *Learner Autonomy and CALL Environments (1st ed.).* Routledge. https://doi.org/10.4324/9780203012208

Settles, B. & Meeder, B. (2016). A trainable spaced repetition model for language learning. *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp.1848–1858). 10.18653/v1/P16-1174.

Sinclair, J. (1991). *Corpus, concordance, collocation: Describing English Language*. Oxford University Press.

Szudarski, P. (2017). *Corpus linguistics for vocabulary: A guide for research (1st ed.).* Routledge. https://doi.org/10.4324/9781315107769

The CEFR-J Wordlist Version 1.6. Compiled by Yukio Tono, Tokyo University of Foreign Studies. Retrieved from http://www.cefr-j.org/download.html#cefrj\_wordlist on 12/12/22.

Thomson, R., & Mehring, J. (2016). Better vocabulary study strategies for long-term learning. *Kwansei Gakuin University Humanities Review, 20,* 133–141.

Thornbury, S. (2002). *How to teach vocabulary.* Pearson Education ESL.

Tono, Y. (2022, October 16). *CEFR-J Reference Level Descriptions (RLDs)*. 東京外国語大学 Tokyo University

of Foreign Studies. Retrieved December 23, 2022, from https://www.tufs.ac.jp/ts/personal/corpuskun/wiki/index.php?CEFR-J%20RLD

Uchida, S. (2021). CVLA: CEFR-based Vocabulary Level Analyzer (ver. 2.0) [Web application]. https://cvla.langedu.jp/

Wilkins, D. A. (1972). *Linguistics in language teaching*. Edward Arnold.

Yamamoto, Y. (2014). Multidimensional vocabulary acquisition through deliberate vocabulary list learning. *System*, *42*, 232–243.