APPLYING OPTICAL CHARACTER RECOGNITION AND DEEP LEARNING TO A WEB APPLICATON FOR STUDYING ENGLISH WORDS IN A SPECIALIZED FIELD OF STUDY

S. Okamoto[a], Y. Ashida [b], M. K. Higa* [c], M. Morita [d] and T. Kawakami [e]

[a] National Institute of Technology, Matsue College/Advanced Electronic and Information Systems Course, Shimane, Japan
[b] National Institute of Technology, Matsue College/Department of Electrical and Computer Engineering, Shimane, Japan
[c] Hiroshima University/Institute for Foreign Language Research and Education, Hiroshima, Japan
[d] Hiroshima University/Department of Physics, Faculty of Science, Hiroshima, Japan
[e] Hiroshima University/Advanced Science and Engineering, Hiroshima, Japan

M. K. Higa* (mhiga@hiroshima-u.ac.jp)

Many English vocabulary learning applications are oriented towards general English contents and are not be suitable for university and vocational students to learning English vocabulary in their specialized fields. Therefore, we have developed Hi-lex, an English learning web application aimed at students learning English vocabulary in specialized fields. Hi-lex was developed using Python and Django, a web application framework for Python. In this study, we implemented a sentence input function using Optical Character Recognition (OCR) and an article suggestion function that enables users to learn by suggesting articles related to the Hi-lex helps users improve their English language skills by allowing them to create and store a customized vocabulary list from the English sentences they want to study. Users can then review their stored words in a flashcard-style format with their desired device. Since its launch in September 2023, the only way to enter text into Hi-lex was by copy and pasting or directly entering text into a the program. To enable learners to learn English words from printed texts, photographs and uneditable computer files, we added a text input function using OCR. To implement this function, we used open-source Python libraries such as Pytesseract, PyOCR, Pillow, and pdf2image. The implemented function first identifies whether the file is a PDF file or not when the file is entered into the file form, and converts it to an image file using Pdf2image if it is a PDF file. Then, it extracts the English words in the image using PyOCR and Pytesseract. The extracted English words are tokenized by the Natural Language Toolkit (NLTK), a Python plugin that was already used by the program. Once this process is finished, a profile of each word is displayed and a student can immediately obtain information such as the definition of the word, it's part of speech, related field and level of difficulty.

Additionally, to make it easier for users to learn English words related to their major field of study, an article compliation function that enables users to learn by compiling articles related to the input text has been implemented in Hi-lex. The function summarizes text input into Hi-lex by user and uses the API to search for and present articles related to the summarized sentences on arXiv.

Keywords: *OCR, Second language acquisition, Spaced repetiton software, Autonomous learning*

Introduction

Vocabulary acquisition is a complex and crucial aspect of second language learning (Dionisio et al., 2022). While learning through applications has proven to be a useful method, most widely available applications focus on general vocabulary, making them less suitable for students in higher education who need to learn specialized terminology within their fields of study. To address this need, researchers at Hiroshima University and the National Institute of Technology, Matsue College have been developing and operating an English learning support web application since September 2023. This application, named the Hi-lex system, is built using Python and a web application frameword, called Django, and aims to provide a practical and efficient learning tool. Hi-lex allows users to find and learn specialized vocabulary, whether unfamiliar or familiar, as described by Fraser et al. (2025). The system features a front end (e.g., HTML, Bootstrap) and a back end (e.g., Python, JavaScript, Django), utilizing word lists and test data stored in a database to facilitate user learning. It is publicly accessible on AWS Lightsail and is currently being used by approximately 130 students at Hiroshima University and National Institute of Technology, Matsue College.

This paper details the development and evaluation of two extended functionalities implemented in Hi-lex: a function for inputting text from files using OCR and a function for searching for related articles based on the input text. The goal of these new features is to further enhance learning opportunities for users.

Usage of the Software

Fig. 1 illustrates the Hi-lex learning process. First, a user inputs English text into the system. The Natural Language Toolkit (NLTK) then transforms the words in the sentence into their basic forms and displays information such as meanings, parts of speech, and difficulty levels. These word lists provide crucial information that helps learners select both common and specialized vocabulary (Higa and Ashida, 2023). NLTK also supports sentence classification, normalization, and parsing. After reviewing the displayed words, the user selects the necessary English words to create a personalized vocabulary list. The system then generates word tests from this created list. To optimize memory retention, the word utilizes the SuperMemo2 algorithm, which determines the optimal interval for reviewing each word. Fig. 2 shows an example of the Hi-lex English vocabulary review function.

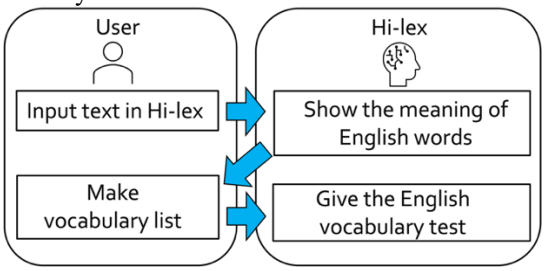


FIGURE 1. The Chart of Hi-lex Learning Process.

FIGURE 2. English Vocabulary Review Function

Development of the Sentence Input Function Using OCR

Previously, the only way to input English text into Hi-lex was for a user to type or paste it into a text field. This made it difficult to handle multi-page documents and handwritten text. To overcome this limitation, we implemented an OCR (Optical Character Recognition) function in Hi-lex. This feature recognizes characters in image data, converts them into text, and can extract English words from PDF and image files. The English input form is shown in Fig. 3. To accommodate PDF and image file uploads, we added a new file form using Python, and toggle buttons were added with HTML, CSS, and JavaScript to switch between the two input methods.

To implement this functionality, we utilized several libraries: Pillow for Python image processing, Pytesseract as the OCR engine, and PyOCR to act as a bridge between Python and the OCR engine.

(a)Text Form (b)File Form

FIGURE 3. The File Input Form

Since OCR primarily supports image files, we also integrated pdf2image to convert PDF files into images before processing. The file input workflow is illustrated in Fig. 4.

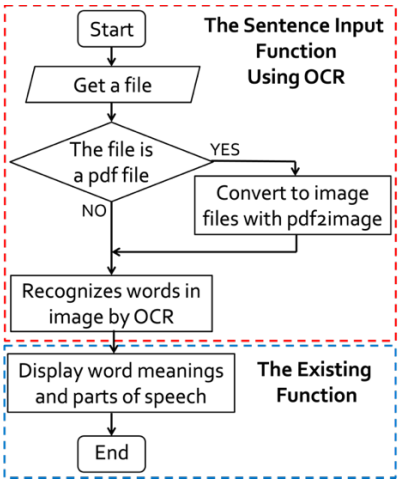


FIGURE 4. File Input Flow

The newly implemented OCR function in Hi-lex first identifies whether the file submitted by the user is a PDF or an image. If an image file is submitted, it uses OCR to recognize the English words within the image and converts them into text data. If a PDF file is submitted, it is first converted into an image file by pdf2image, and the same process is then applied. The resulting text data is then passed to existing functions to display associated meanings, parts of speech, and difficulty levels from the word lists.

An academic article of 326 words, shown in Fig. 5, was used for operational testing. After excluding common words that are not displayed, words not in the Hi-lex word list, duplicates, and proper nouns, the total number of target words was 95. The image file used for testing was a screenshot of the same article, with an image size of 750px × 1060px.

Fig. 6 presents the operational results. The PDF file conversion took 12.7 seconds from input to output and correctly identified 83 words (87.3% accuracy). The image file conversion took 7.9 seconds and correctly identified 80 words (84.2% accuracy). In both cases, the system successfully recognized the English words in the files and displayed their meanings, parts of speech, and

difficulty levels from the word lists. However, a notable limitation was that English words with line breaks or hyphens were not recognized, and some words not present in the original text were displayed due to misrecognition.

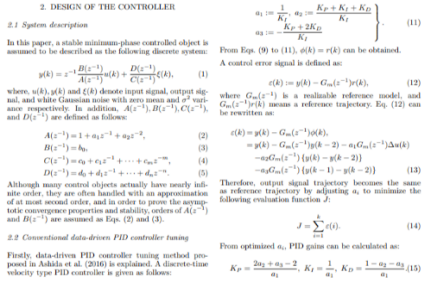
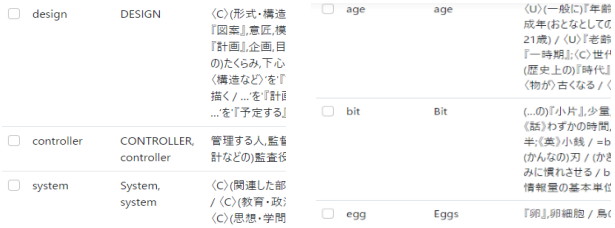


FIGURE 5. Article Used to Check the Operation



(a) Correctly Displayed (b) Misrecognition

FIGURE 6. Operational Results

Development of the Articles Suggestion Function

The current Hi-lex system allows users to learn the general meanings and example sentences of English words related to their field of study by manually entering them into a form. However, it lacks a specific function to facilitate this type of contextual learning. To enhance learning opportunities for users by exposing them to relevant English words and academic articles, we implemented a new function that suggests suitable articles based on user input.

The proposed system uses an LLM (Large Language Model) and aricles Database, a search engine specializing in academic information. Fig. 7 illustrates the workflow of this article suggestion function.

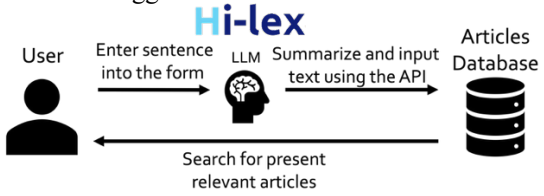


FIGURE 7. Chart of The Input-Friendly Article Presentation Function

A user first enters a sentence into the form, which is then summarized by the LLM. The LLM subsequently uses an API (Application Programming Interface) to send this summarized content to arXiv. It then receives information on relevant articles, which it summarizes and presents to the user in a list. The user can select articles of interest using a checkbox. The selected article is then automatically input into the form, and the system displays the meanings of the English words within it to facilitate further learning.

For summarizing the input text, we used Llama 3, a large language model published by Meta. To run the LLM locally without a network connection, we utilized the llama.cpp open-source platform.

Fig. 8 shows the operational results. Clicking a blue button displays the title, authors, and links of articles that are related to the text entered into the form.

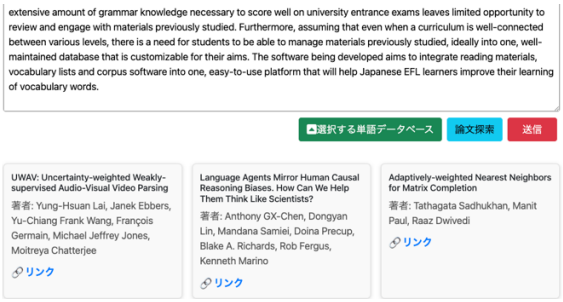


FIGURE 8. Operational Results

Clicking on the link will take you to Arxiv, where user can learn more English words related to specialized field of study. Additional functions, such as paper translation and keyword extraction, can also be implemented.

Conclusions

In this paper, we described the implementation of two new functions for the Hi-lex system: one for text input from files using OCR and another for suggesting relevant academic articles. We confirmed the successful operation of both functions through testing. Moving forward, the plan is to further enhance these functionalities and to ensure server redundancy. These improvements aim to provide a more comfortable and proactive English learning environment for users.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP20K13155.

References

Higa, M. (2025). *Hi-lex System* [Web application]. Retrieved from https://hi-lex.org

Dionsio, G. & Pascual, L. & Ramil G. (2022). Vocabulary and Learning Strategies in Second Language Learning. *International Journal of English Language Studie*, 4(3), 9.

Simon, F. & Higa, M. & Walter, Davies. (2025). Delivering an ESP Pedagogic Word List: Integrating Corpus Analysis, Materials Design, and Software Development. *Languages,* 10(3), 46. https://doi.org/10.3390/languages10030046.

Higa, M. & Ashida, Y. (2023). A Pedagogical Framework for the Development of Software for EFL Learners. *Hiroshima Studies in Language and Language Education,* 26, 63-77.