Asia Pacific Journal of Corpus Research
Vol. 6, No. 2

http://www.icr.or.kr/apjcr
e-ISSN 2733-8096
https://doi.org/10.22925/apjcr.2025.6.2.1

VACSR Version 4: Enhancing a CEFR-J-Based Vocabulary Self-Reflection Tool for Multilingual and Foreign Language Education

Yukiko Ohashi
(Yamazaki University of Animal Health Technology)

Ohashi, Y. (2025). VACSR version 4: Enhancing a CEFR-J-Based vocabulary self-reflection tool for multilingual and foreign language education. *Asia Pacific Journal of Corpus Research, 6*(2), 1-14.

This study introduces VACSR v.4.0, an updated version of the Vocabulary Analyzer for Self-Reflection. The tool automatically analyzes vocabulary occurrences and CEFR-J levels in transcribed texts, providing information on usage frequency, unused items, and tokens not classified within CEFR-J scales. VACSR v.4.0 incorporates syntactic parsing using the Stanza library, enabling the system to distinguish parts of speech and categorize identical lexical items according to their grammatical functions. The updated version also adds the automatic display of top headwords, a feature not included in previous versions, and extends its processing capability to tokens containing diacritical marks, such as ê and û, found in other languages. The study aims to explore how VACSR v.4.0 functions with languages other than English through a pilot analysis using two French texts containing vocabulary with diacritical marks. Although the system does not yet provide linguistically accurate POS tagging or CEFR-based level assignments for French, it can still extract high-frequency vocabulary and generate meaningful frequency-based lists. These findings show that, even without full multilingual NLP integration, VACSR v.4.0 helps compile core vocabulary lists across different languages and offers preliminary insights for pedagogical decision-making in foreign-language learning contexts.

Keywords: Corpus Linguistics, French Vocabulary, CEFR-J, Vocabulary, Language Education

1. Introduction

Vocabulary acquisition is a cornerstone of language learning, playing a crucial role in learners' ability to communicate effectively across various contexts. The Common European Framework of Reference for Languages (CEFR) serves as a comprehensive guideline for language proficiency, outlining levels from A1 (beginner) to C2 (proficient). Building on the CEFR, the CEFR-J was developed specifically to address the needs of Japanese learners of English, offering a finer-grained categorization of vocabulary and language competencies at lower proficiency levels (Tono, 2013).

Research into vocabulary acquisition has highlighted the importance of frequency, context, and repetition in facilitating retention and usage (Nation, 2001, 2013; Schmitt, 2010). Studies such as Milton (2009) emphasize the relationship between vocabulary size and language proficiency, demonstrating that learners with a broader vocabulary range tend to perform better in communicative tasks. Similarly, Laufer and Nation (1995) proposed the "Lexical Threshold Hypothesis", which argues that a minimum vocabulary size is essential for adequate reading comprehension.

Among Japanese learners, previous studies have shown that vocabulary instruction often emphasizes rote memorization and test preparation, thereby neglecting the development of more profound lexical knowledge and communicative competence (Takahashi, 2015). The CEFR-J aims to bridge this gap by providing a structured framework for vocabulary teaching that aligns with learners' progressive development.

Recent advancements in corpus linguistics and vocabulary assessment tools have facilitated data-driven approaches to vocabulary instruction. Tools such as the Classroom Corpus Vocabulary Analyzer with the CEFR-J Wordlist (CCVA; Ohashi et al., 2021) and the Vocabulary Analyzer based on CEFR-J Wordlist for Self-Reflection (VACSR; Ohashi & Katagiri, 2022; Ohashi et al., 2023) allow teachers and researchers to examine vocabulary distribution in classroom corpora and reflect on pedagogical practices. VACSR, in particular, has demonstrated practical utility in English language education by enabling users to analyze vocabulary levels and frequency distributions using the CEFR-J framework.

While VACSR has been effectively used in English language contexts, its applicability to languages other than English remains largely unexplored. Investigating this question is increasingly relevant as multilingual learning environments expand and educators seek tools that can support vocabulary analysis across languages.

However, because the CEFR-J is an English-specific lexical resource, any claims regarding the multilingual applicability of VACSR must be made with caution. The current version cannot provide linguistically accurate part-of-speech (POS) classification or CEFR-based proficiency assignment for languages other than English. To explore how VACSR v.4.0 might nevertheless be used with texts written in languages other than English—and to identify both the potential educational benefits and the limitations involved—this study uses French texts as a test case. By doing so, we aim to clarify the extent to which VACSR can support vocabulary-related analyses in a non-English context and the considerations necessary when applying the tool beyond its original design. Accordingly, the present study formulates the following research questions:

RQ1. To what extent is VACSR v.4.0 technically capable of processing non-English texts when language-specific CEFR-aligned resources are not yet implemented?

RQ2. What kinds of frequency-based vocabulary lists does VACSR v.4.0 generate across multiple French texts, and to what extent can these outputs inform pedagogical use while acknowledging system constraints?

2 Background

2.1. The CEFR-J Framework: An Overview

The CEFR-J (Common European Framework of Reference for Languages: Japan-specific Framework) is an adaptation of the CEFR, tailored specifically to the context of English education in Japan. Unlike the CEFR, which categorizes language proficiency into six levels (A1, A2, B1, B2, C1, and C2), the CEFR-J further subdivides these levels. It introduces finer gradations—pre-A1, A1.1 to A1.3, A2.1 to A2.2, B1.1 to B1.2, B2.1 to B2.2, C1, and C2—allowing for more precise evaluation of learners' abilities at the lower proficiency stages.

In addition to these sublevels, the CEFR-J provides detailed descriptors across five communicative skills: listening, reading, speaking in interaction, speaking in production, and writing. These descriptors are designed to reflect the learning patterns of Japanese English learners, the majority of whom fall within the A1 proficiency band (Negishi et al., 2013). The classroom corpora analyzed by Ohashi and Katagiri (2020) also demonstrated that words classified at the A1 level were frequently utilized in Japanese elementary schools. Considering these circumstances, it is essential to develop instructional guidelines based on the CEFR-J levels that incorporate vocabulary instruction informed by current situation analyses to facilitate learners' progression beyond the A1 level.

2.2. The CEFR-J Wordlist

The CEFR-J Wordlist was constructed using corpora of English textbooks widely used across primary to secondary schools (Years 3 to 10) in countries such as China, Korea, and Taiwan (Tono, 2013). This approach ensures that the vocabulary reflects learners' actual lexical needs in the region. The word list is organized by level, and some words may appear on multiple levels depending on their parts of speech. The CEFR-J Wordlist categorizes a total of 7,799 vocabulary items across four primary proficiency levels: A1, A2, B1, and B2. At the A1 level, which ranges from beginner to lower-intermediate proficiency, there are 1,164 words. The A2 level, representing upper-beginner to lower-intermediate learners, includes 1,411 words. For learners at the B1 level, which spans lower-intermediate to intermediate proficiency, the wordlist contains 2,446 items. Finally, the B2 level, corresponding to upper-intermediate to advanced learners, encompasses 2,778 words. By applying the CEFR-J framework to classroom corpora, educators can identify which vocabulary levels are most prevalent in specific learning environments or determine which vocabulary items should be introduced to align with curriculum goals.

The integration of classroom corpora provides language teachers with valuable insights into their teaching performance, particularly in terms of CEFR-J wordlist-aligned vocabulary usage and the balance of utterances between teachers and students. To facilitate such reflection, Ohashi and Katagiri. (2022) developed VACSR v.1.0, a tool designed to analyze the frequency and proficiency levels of vocabulary items used in classroom settings. By enabling teachers to identify patterns in their language use, VACSR v.1.0 helps promote a more informed approach to teaching. Additionally, the tool supports simultaneous analysis of multiple texts, allowing comparisons of word frequencies across files based on vocabulary items listed in the CEFR-J wordlist.

Building on these capabilities, the revised version, VACSR v.2.0 (Ohashi et al., 2023), addressed several limitations of its predecessor. Notable improvements include the ability to (1) distinguish words with identical spelling but different parts of speech (e.g., "mean" as a verb [A1], adjective [A2], and noun [B1]) and (2) recognize multi-word tokens (MWTs), such as "well-known". VACSR has undergone several iterations since its initial release.

While VACSR v.3.0 was developed internally, it involved only minor adjustments to VACSR v.2.0 and was not released. The present paper, therefore, introduces VACSR v.4.0, which includes several substantial improvements. VACSR v.4.0 enhances usability and operational efficiency through a more streamlined interface and improved processing workflow. Although the tool remains primarily designed for English and the CEFR-J framework, its expanded handling of tokens containing diacritical marks provides the basis for examining how the system processes vocabulary in languages that differ orthographically from English.

3. Method

3.1. VACSR v.4.0 System Configuration

This section outlines the operation of VACSR v.4.0, which is currently available at https://cctvtt.com/vacsr4/. To ensure methodological transparency and replicability, the computational configuration of VACSR v.4.0 is described in detail below. VACSR v.4.0 was developed using Stanza (Qi et al., 2020) and applies the English EWT (English Web Treebank) model for part-of-speech (POS) tagging. Within VACSR v.4.0, Stanza is called with its default English EWT pipeline, which includes tokenization, multi-word token (MWT) expansion, and POS tagging. VACSR v. 4.0 employs the same methodology as its predecessor, VACSR v.2.0 (Ohashi et al., 2023). This approach ensures consistency in parsing functionality, particularly through Stanza's part-of-speech (POS) tagging system, which classifies words into 36 categories. The conversion method between Stanza's POS tags and the CEFR-J wordlist was retained, effectively resolving tag classification discrepancies and maintaining compatibility for text analysis. No language-specific POS model for French was implemented in the current study, as VACSR v.4.0 is designed around CEFR-J, which is an English-specific lexical resource.

3.2. POS-to-CEFR-J Mapping

VACSR v.4.0 assigns CEFR-J levels (A1–B2) to all tokens that appear in the CEFR-J wordlist, regardless of whether they are content words or function words. Thus, articles, pronouns, conjunctions, and other grammatical items also receive CEFR-J levels when listed in the CEFR-J wordlist.

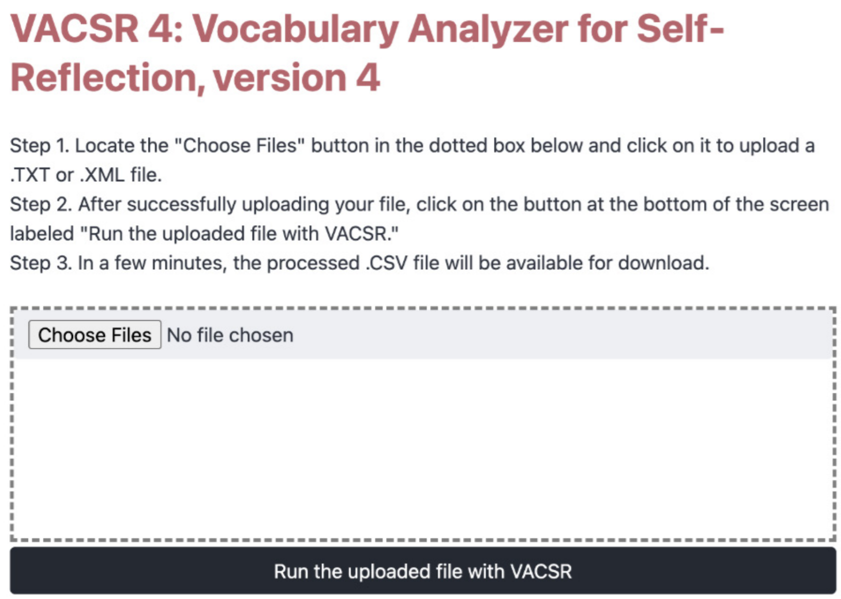
Tokens not included in the CEFR-J wordlist—including proper nouns, loanwords, and other out-of-list items—are classified into the “other” category. Because VACSR v.4.0 is designed for English and relies on an English-specific CEFR-J list, non-English vocabulary items are also assigned to “other,” except for cases in which their surface forms coincide with English words. Because the CEFR-J is English-specific, French tokens either (a) receive English-based CEFR-J levels when they share spelling with English words (e.g., car, or, ton), or (b) are placed in the other category when no orthographic match exists.

Tokenization and normalization in VACSR v.4.0 follow the default settings of Stanza’s English model. All alphabetic characters are converted to lowercase, and punctuation marks are systematically separated from adjacent word tokens. Stanza also divides clitic contractions into their component parts (for example, don’t is split into do and n’t), enabling consistent token-level analysis. Importantly, VACSR preserves diacritics in the original text and does not apply any form of diacritic normalization; thus, French words such as école remain in their original orthographic form. In addition, the system does not perform lemmatization, stemming, or any morphological normalization, meaning that inflected word forms (e.g., parle, parles, parler) are treated as separate types. This configuration reflects the system’s design as a surface-level vocabulary profiling tool and ensures that all word forms appearing in the input text are preserved in the output.

VACSR allows users to analyze multiple corpora simultaneously by comparing word occurrences and CEFR-J–based vocabulary levels across files. As in previous versions, all extracted vocabulary items are categorized into four CEFR-J levels (A1–B2). Figure 1 shows the initial interface, and the analysis procedure follows the same three-step workflow used in VACSR v.2.0.

3.3. Procedure of Using VACSR v. 4.0

To begin using VACSR v.4.0, Step 1 involves uploading a file in .txt or .xml format by clicking the “Choose Files” button in Figure 1. Once the file is successfully uploaded, Step 2 requires users to initiate the analysis by clicking the “Run the uploaded file with VACSR” button. Step 3 entails the program executing the analysis, which may take a few minutes. During this time, users are advised to refresh the browser periodically. Once the execution is complete, the result file will be saved in the designated directory. It is important to note that the generated result files will remain accessible for only three hours after creation.



VACSR 4: Vocabulary Analyzer for Self-Reflection, version 4

Step 1. Locate the "Choose Files" button in the dotted box below and click on it to upload a .TXT or .XML file.
Step 2. After successfully uploading your file, click on the button at the bottom of the screen labeled "Run the uploaded file with VACSR."
Step 3. In a few minutes, the processed .CSV file will be available for download.

Figure 1. VACSR v.4.0. Viewing Screen

3.4. Improvements in VACSR 4.0

Building on the functionality of VACSR v.2.0, VACSR v.4.0 retains the ability to calculate word frequencies based on their alignment with CEFR-J vocabulary levels (A1, A2, B1, B2) across multiple uploaded texts. This functionality provides a clear overview of the vocabulary distribution within an analyzed text. Specifically, the tool identifies which words from the CEFR-J levels appear in the text and how frequently they are used, while also highlighting words from each level that are absent. Words that are present in the text are displayed with their respective frequency counts under each CEFR-J level (as exemplified in Table 1). In contrast, words that do not appear in the text are enclosed in parentheses—(A1), (A2), (B1), or (B2)—and are listed under their respective CEFR-J levels with a frequency of zero (as illustrated in Table 2). This structured output provides an immediate, comprehensive understanding of both the vocabulary level and the frequency of each word in the analyzed text. As with VACSR v.2.0, VACSR v.4.0 also quantifies the distribution of identical words across different texts using a RANGE value. For instance, if a word appears in both of the two analyzed texts, its RANGE value is recorded as 2, whereas if it occurs in only one text, the RANGE value is 1.

Notable advancements in VACSR v.4.0 include the following:

(1) the implementation of a new feature that automatically generates frequency-based vocabulary lists as “Top Headwords” by ranking words observed at each RANGE level in descending order of frequency across all analyzed texts; and

(2) improved handling of tokens containing diacritical marks (e.g., é, è, ê, à, â, î, ù, û, ô, ç, ë, ï), allowing such items to be displayed correctly in the output. As VACSR relies on the CEFR-J wordlist, vocabulary items not included in the A1–B2 levels—including most non-English words—are categorized as “other.”

Tables 3 and 4 present partial excerpts from the vocabulary lists automatically generated by VACSR v.4.0. Table 3 illustrates the Top Headwords list for the English texts, showing the most frequent lexical items shared across the two texts, along with their POS categories, CEFR-J levels, RANGE values, and token frequencies. High-frequency function words (e.g., you, so, and, I) as well as common verbs (e.g., see, think, take) appear prominently, demonstrating VACSR’s ability to extract shared high-frequency vocabulary across multiple English texts.

Table 4, in contrast, presents French vocabulary items that were classified as “other” when Corpus

1 and Corpus 2 were processed. These entries include many words with French-specific diacritical marks and inflected forms (e.g., ça, échanger, école, économique), which do not appear in the CEFR-J wordlist. Because VACSR assigns all tokens not found in CEFR-J levels A1–B2 to the "other" category, French words whose spellings do not match English items are placed in this group. Since these French-specific items are not included in English lexical resources, their parts of speech cannot be determined by the English POS model and therefore receive the label "X." In addition, it is important to clarify why certain French words receive the "X" label while others—particularly French–English homographs—do not. VACSR v.4.0 relies on Stanza's English POS tagging model, and no French-specific model is currently implemented. Consequently, French tokens that do not exist in English dictionaries are treated as out-of-vocabulary items and assigned the tag "X." Conversely, French words that share surface forms with English words are recognized by the English POS tagger and thus appear in English POS categories and may even receive CEFR-J levels. This behavior reflects the English-centric architecture of the current system and represents a methodological limitation rather than a malfunction of VACSR.

Despite these constraints, the accurate extraction of French-specific vocabulary into the "other" category demonstrates that VACSR can still generate meaningful frequency-based vocabulary lists for non-English texts. The system's ability to identify high-frequency items and shared vocabulary across texts highlights its potential applicability as a preliminary vocabulary analysis tool in diverse linguistic contexts.

Table 1. Examples of Vocabulary from CEFR-J Wordlist Levels Found in Texts, Extracted from Updated Version, VACSR v.4.0

Headword	POS	CEFR-J	RANGE	FREQ	Text 1	Text 2
radio	noun	A1	1	1	1	0
rat	noun	A1	1	1	1	0
site	noun	A1	2	2	1	1
son	noun	A1	2	30	21	9
sport	noun	A1	2	3	2	1
super	adjective	A1	1	1	0	1
tennis	noun	A1	1	4	4	0
accident	noun	A2	1	1	1	0
association	noun	A2	2	5	2	3
attention	noun	A2	1	1	1	0
bonus	noun	A2	1	1	0	1
compose	verb	B1	2	2	1	1
conclusion	noun	B1	1	1	0	1
courage	noun	B1	1	1	0	1
discrimination	noun	B1	1	1	1	0
distance	noun	B1	2	2	1	1
convention	noun	B2	1	1	0	1
correspond	verb	B2	1	1	1	0
dose	noun	B2	1	1	0	1
exception	noun	B2	1	1	1	0
explorer	noun	B2	1	2	0	2

Table 2. Examples of Vocabulary Not Found in Texts from CEFR-J Wordlist Levels, Extracted from Updated Version, VACSR v.4.0

Headword	POS	CEFR-J	RANGE	FREQ	Text 1	Text 2
almost	adverb	(A1)	0	0	0	0
alone	adverb	(A1)	0	0	0	0
along	adverb	(A1)	0	0	0	0
already	adverb	(A1)	0	0	0	0
advanced	adjective	(A2)	0	0	0	0
advantage	noun	(A2)	0	0	0	0
adventure	noun	(A2)	0	0	0	0
advertisement	noun	(A2)	0	0	0	0
assign	verb	(B1)	0	0	0	0
assignment	noun	(B1)	0	0	0	0
assist	verb	(B1)	0	0	0	0
assistance	noun	(B1)	0	0	0	0
abandoned	adjective	(B2)	0	0	0	0
abnormally	adverb	(B2)	0	0	0	0
abolish	verb	(B2)	0	0	0	0
aboriginal	adjective	(B2)	0	0	0	0

Table 3. Partial Excerpt from the Top Headwords List Automatically Generated by VACSR v.4.0

Headword	POS	CEFR-J	RANGE	FREQ	Text 1	Text 2
you	pronoun	A1	2	225	162	63
so	adverb	A2	2	185	133	52
and	conjunction	A1	2	113	77	36
I	pronoun	A1	2	96	78	18
see	verb	A1	2	57	44	13
think	verb	A1	2	40	35	5
recommend	verb	B1	2	28	14	14
let	verb	A1	2	27	22	5
take	verb	A1	2	27	21	6

Table 4. Partial Excerpt from the French Vocabulary Classified as ‘Other’ by VACSR v.4.0, Including Specialized Terms Generated by VACSR v.4.0

Headword	POS	CEFR-J	RANGE	FREQ	Text 1	Text 2
ça	X	(other)	2	24	14	10
échanger	X	(other)	1	1	1	0
échanges	X	(other)	1	3	0	3
échappé	X	(other)	1	1	1	0
échelle	X	(other)	1	1	0	1
échoué	X	(other)	1	1	0	1
école	X	(other)	1	2	2	0
écoles	X	(other)	2	6	2	4
économique	X	(other)	1	1	0	1
écouter	X	(other)	1	1	1	0
écrire	X	(other)	2	3	1	2

3.5. Pilot Study

This section presents a pilot study using VACSR v.4.0 to examine how the tool behaves when processing texts written in French. The analysis illustrates the extent to which VACSR can generate useful frequency-based vocabulary information in a non-English context and discusses the pedagogical implications of this output. By providing word-frequency data across multiple texts, VACSR v.4.0 offers a practical means of identifying commonly occurring lexical items, a key consideration in vocabulary teaching.

In French language education, corpus-based vocabulary resources such as FLELex (Français Langue Étrangère Lexique; Pintard & François, 2020) are publicly available. While such established lexicons serve as important reference materials, VACSR v.4.0 enables educators and researchers to generate text-specific or domain-focused vocabulary lists directly from their own materials. The capacity to identify high-frequency items shared across multiple texts adds flexibility and enhances its usefulness for instructional planning.

This pilot study, therefore, aimed to explore the potential of VACSR v.4.0 for French as a Foreign Language (FFL). Specifically, it examined the vocabulary shared across two French texts to determine what kinds of frequency-based lists the tool can produce under current system constraints.

3.5.1. Study Materials

Two French-language texts (Corpus 1 and Corpus 2) were selected for the analysis. Both texts address general contemporary topics such as environmental issues, public health, and social challenges—topics that are commonly encountered in intermediate-level FFL (French as a Foreign Language) classrooms. The texts were obtained from *1jour1actu* (https://www.1jour1actu.com/), an educational news website in France designed for children. The site provides freely accessible written and video-based materials created for educational purposes. The specific articles used in the present analysis were downloaded, and only the textual content was used for non-commercial academic research.

Corpus 1 (French 1.txt) consists of 6,476 words, and Corpus 2 (French 2.txt) contains 6,597 words, resulting in a combined corpus of 13,073 words. The two texts were chosen to ensure thematic similarity and a comparable lexical register, making them suitable for identifying shared high-utility vocabulary items.

3.5.2. Study Procedure

The analysis was conducted using VACSR v.4.0, a tool initially developed for English vocabulary analysis. Although VACSR v.4.0 can process French text files, its output still relies on English-specific resources, so most French items are not assigned CEFR-J levels. Nevertheless, the frequency information generated by the system provides valuable preliminary insights for vocabulary-focused analysis.

Two French texts were uploaded into VACSR v.4.0. The tool automatically calculated word frequencies for each file and, when both were analyzed together, identified vocabulary items that appeared in both texts using its RANGE function. These shared items were compiled into a single list and ranked according to their combined frequencies, forming the basis for subsequent qualitative interpretation.

To maintain the scope of this pilot study, only two thematically comparable French texts (Corpus 1 and Corpus 2) were selected. Using a small, genre-consistent dataset enabled a controlled observation of how VACSR's RANGE function behaves across files without the additional variability introduced by larger or more heterogeneous corpora.

Within this focused dataset, verbs and adjectives were chosen for closer examination. These categories play central roles in French vocabulary development and exhibit substantial morphological variation: verbs show extensive inflectional patterns, while adjectives vary in gender and number. By focusing on these two categories, the analysis illustrates how VACSR v.4.0 handles morphologically variable items in a non-English context while maintaining a clear, pedagogically meaningful scope for this preliminary investigation.

In preprocessing, all texts were lowercased, diacritics were preserved as written in the original French text, and contracted forms (e.g., c'est, du, au) were not de-contracted. No additional normalization procedures were applied.

4. Results

A total of 746 wordforms were identified in the RANGE 2 list, representing all tokens that appeared in both French texts. RANGE is defined as the number of texts in which a given headword appears. Because the pilot dataset consists of two texts, RANGE takes a value from 0 to 2.

To focus the analysis on lexically meaningful items, all wordforms occurring only once across the two texts (RANGE 1) were removed, as the items appearing only once in the entire dataset provide little pedagogical value in a pilot aimed at identifying pedagogically salient vocabulary. After excluding these hapax legomena, 402 shared wordforms remained. From this reduced set, verbs and adjectives were extracted for closer analysis. For verbs, all inflected surface forms belonging to the same lexical item (e.g., parle, parles, parlons, parlé) were consolidated and counted as a single lemma. Adjectives were treated in the same way: gender and number variants (e.g., nouveau, nouvelle, nouveaux, nouvelles) were merged and counted as a single lexical item. After this consolidation process by hand, 22 verb lemmas (Table 5) and 23 adjective lemmas (Table 6) were identified, and all of these items were included in the final lists for analysis.

Because VACSR v.4.0 relies on an English POS-tagging model, most French vocabulary items were assigned the POS label "X," and French words whose spellings overlap with English occasionally received incorrect English POS tags or CEFR-J levels (e.g., important; see Table 6). For this reason, the automatically generated POS and CEFR-J information was not used when compiling the verb and adjective lists. Instead, VACSR v.4.0 was used only to extract surface wordforms and their frequencies across the two texts. Since the system outputs inflected forms separately, all vocabulary items were manually checked, and morphologically related forms were consolidated into lemma-level entries based on French dictionary criteria. The resulting lemma-based lists (Tables 5 and 6) provide a more precise representation of high-frequency French vocabulary shared across the texts and offer pedagogically valuable insights into items that are likely to be contextually versatile. Although this pilot study focuses on French, the same lemma-consolidation procedure can be applied to other European languages, demonstrating how VACSR v.4.0 can support frequency-based vocabulary selection even when complete multilingual NLP resources are not yet integrated.

Table 5. List of the 22 Consolidated Verb Lemmas Shared by Corpus 1 and Corpus 2 (Identified by VACSR v.4.0)

Headword	POS	CEFR-J	RANGE	Frequency	French1.txt	French2.txt
être	X	(other)	2	390	191	199
faire	X	(other)	2	54	31	23
manger	X	(other)	2	15	10	5
savoir	X	(other)	2	14	8	6
aider	X	(other)	2	21	12	9
donner	X	(other)	2	9	5	4
demander	X	(other)	2	12	5	7
protéger	X	(other)	2	9	5	4
changer	X	(other)	2	12	5	7
mettre	X	(other)	2	9	5	4
passer	X	(other)	2	6	5	1

respecter	X	(other)	2	6	4	2
comprendre	X	(other)	2	5	3	2
menacer	X	(other)	2	7	4	3
partir	X	(other)	2	5	3	2
prendre	X	(other)	2	8	4	4
vouloir	X	(other)	2	9	4	5
aller	X	(other)	2	16	6	10
obliger	X	(other)	2	4	2	2
permettre	X	(other)	2	4	2	2
viennent	X	(other)	2	15	4	11
vérifier	X	(other)	2	4	2	2

Table 6. List of the 23 Consolidated Adjective Lemmas Shared by Corpus 1 and Corpus 2 (Identified by VACSR v.4.0)

Headword	POS	CEFR-J	RANGE	Frequency	French1.txt	French2.txt
cette	X	(other)	2	56	28	28
certains	X	(other)	2	19	9	10
chaque	X	(other)	2	15	11	4
mondiale	X	(other)	2	15	8	7
plastique	X	(other)	2	15	2	13
internationale	X	(other)	2	13	4	9
autres	X	(other)	2	12	3	9
bon	X	(other)	2	12	3	9
plusieurs	X	(other)	2	12	8	4
nombreux	X	(other)	2	11	7	4
politique	X	(other)	2	10	5	5
scientifique	X	(other)	2	10	4	6
important	adjective	A1	2	9	7	2
petit	X	(other)	2	9	6	3
nouveau/nouvelle	X	(other)	2	8	5	6
pauvres	X	(other)	2	8	3	5
célèbre	X	(other)	2	7	3	4
climatique	X	(other)	2	6	3	3
grande	X	(other)	2	6	3	3
difficile	X	(other)	2	5	3	2
seul	X	(other)	2	5	3	2
forte	X	(other)	2	4	1	3
populaire	X	(other)	2	4	2	2

5. Discussion

This section discusses the study’s findings by addressing the two research questions. We first examine the technical behavior and limitations of VACSR v.4.0 when processing non-English texts in the absence of language-specific CEFR-aligned resources (RQ1). We then consider the types of frequency-based vocabulary lists the system generates across the two French texts and the extent to which these outputs can inform pedagogical use under the current system constraints (RQ2).

5.1. Answers to Research Questions

Answer to RQ1.

The application of VACSR v.4.0 to French texts revealed clear patterns regarding its technical capabilities for processing non-English input. The system successfully extracted raw frequency information for French vocabulary items, including those with diacritics, and consistently grouped them into the "other" category in the output. This enabled VACSR to generate basic frequency-based vocabulary lists and shared RANGE lists across the two French texts, indicating that the tool can function as a preliminary vocabulary profiling system even without language-specific CEFR resources.

At the same time, the analysis highlighted the limitations of this English-based architecture. Vocabulary items whose spellings overlap with English (e.g., important) were incorrectly assigned English POS tags or CEFR-J levels, while French-specific forms received the default POS label "X." These outcomes reflect the current design of VACSR v.4.0, which applies English POS tagging and CEFR-J mappings to all tokens regardless of language. As a result, POS and proficiency-level information cannot be interpreted meaningfully for non-English data. Overall, the behavior of VACSR v.4.0 suggests that the tool can provide meaningful frequency-based insights for French texts, while highlighting the need for language-specific tagging and lexicons for more linguistically precise analyses.

Answer to RQ2.

VACSR v.4.0 generated several types of frequency-based vocabulary lists across the two French texts, including (a) surface-form frequency lists, (b) verb and adjective families in which inflectional variants were manually consolidated into lemma-level entries, and (c) RANGE-based headwords representing items occurring in both texts. Together, these outputs provide an initial overview of lexical concentration, shared vocabulary, and potential high-frequency pedagogical targets within the dataset. For example, the 402 items with a RANGE value of 2 represent the core vocabulary shared across the texts and may serve as a starting point for developing glossaries or pre-reading activities.

Interpretation of these lists, however, must take into account the current system constraints. Because VACSR v.4.0 relies on an English-trained POS tagger and CEFR-J mapping, French items were frequently assigned the POS label "X," and English-based POS or CEFR-J levels were occasionally applied to French words with overlapping spellings. In addition, lemma-based aggregation was performed manually and remains limited. Consequently, the lists produced here should be regarded as exploratory indicators rather than definitive pedagogical recommendations.

Looking ahead, incorporating a French POS-tagging model and language-specific CEFR-aligned lexical resources will be required to improve the linguistic accuracy and pedagogical applicability of VACSR's frequency-based outputs in multilingual contexts. Despite these limitations, the findings indicate that VACSR v.4.0 can still serve as a practical tool for generating preliminary frequency-based vocabulary lists in languages other than English, particularly when the instructional focus is on identifying core lexical items rather than conducting detailed linguistic analysis.

5.2. Limitations

Although VACSR v.4.0 was initially designed for English, the present study demonstrates that the tool can extract frequency information from texts written in other languages. As discussed in the preceding sections, several limitations must be acknowledged when interpreting the results of the French pilot analysis.

First, both the CEFR-J wordlist and the POS-to-CEFR mapping embedded in VACSR v.4.0 are English-specific resources. Because the system relies on Stanza's English POS-tagging model, French tokens that do not appear in the CEFR-J are automatically categorized as "other." At the same time, items whose spellings overlap with English may be incorrectly assigned English POS tags or CEFR-J levels.

These English-driven misclassifications are illustrated in Table 7, which presents examples of French words with English-identical spellings that appeared in Corpus 1 and Corpus 2. Under the current design, VACSR cannot provide linguistically valid POS or CEFR-based information for French or other non-English languages.

Second, VACSR v.4.0 does not yet incorporate language-specific NLP pipelines such as Stanza's French models or French CEFR-aligned lexical resources (e.g., FLELex). Integrating such components would require substantial system modification and lies beyond the scope of this exploratory pilot. Full multilingual functionality will require language-specific taggers, lexicons, and CEFR-aligned resources.

Third, the absence of a built-in lemmatization function limits the interpretability of the vocabulary lists. All surface forms are treated as separate types, resulting in fragmented frequency counts, especially for morphologically rich languages such as French. In this study, inflected forms were manually consolidated into lemma-level entries for the purpose of analysis, but automatic lemmatization will be essential for future versions of VACSR to support more accurate vocabulary profiling.

Overall, VACSR v.4.0 is not yet sufficient for complete linguistic analysis of French; however, its ability to extract raw frequency information and identify vocabulary shared across texts indicates that it can still serve as a useful preliminary tool for generating frequency-based vocabulary lists in languages other than English. Future development will incorporate UD-compliant language-specific POS models, CEFR-aligned lexical resources, and automatic lemmatization to enable more accurate multilingual vocabulary analysis.

Table 7. Partial Excerpt of French Vocabulary Items with Spellings Identical to Their English Counterparts, Generated by VACSR v.4.0

Headword	POS	CEFR-J	RANGE	FREQ	Text 1	Text 2
tennis	noun	A1	1	4	4	0
population	noun	A2	1	4	0	4
pirate	noun	B2	1	4	0	4
but	conjunction	A1	1	3	0	3
date	noun	A1	1	3	3	0
condition	noun	A2	1	3	0	3
situation	noun	A2	1	3	0	3
boom	noun	B1	1	3	0	3
client	noun	B2	1	3	3	0

6. Conclusion

This study presented VACSR v.4.0, an enhanced vocabulary analysis tool capable of generating frequency-based wordlists and identifying shared lexical items across multiple texts. Although initially developed for English, the pilot analysis demonstrated that VACSR can extract reliable frequency information from non-English texts as well. Using two French texts as test cases, the tool successfully identified shared high-frequency vocabulary and compiled verb and adjective lists, offering insight into core lexical items that may support pedagogical decision-making.

At the same time, the French analysis also highlighted explicit system constraints. Because VACSR v.4.0 currently relies on an English POS-tagging model and CEFR-J mapping, the tool cannot yet provide linguistically accurate POS categories or CEFR-aligned level information for French. Furthermore, the absence of automatic lemmatization requires manual consolidation of inflected forms when generating pedagogically meaningful vocabulary lists. These findings underscore that VACSR's present multilingual functionality remains exploratory and primarily limited to raw frequency extraction.

Nevertheless, within these constraints, VACSR v.4.0 can still serve as a practical preliminary tool for compiling frequency-based vocabulary lists in languages other than English, particularly when instructors seek to identify core lexical items shared across multiple texts. Future development will focus on integrating language-specific POS taggers, CEFR-aligned lexical resources, automatic lemmatization, and larger corpora to support more linguistically precise and pedagogically robust applications across a broader range of languages.

Acknowledgments

This study is part of a research program financed by the Promotion of Science (JSPP), Grant-in-Aid for Scientific Research (C) No. 23K00708

Code and Data Availability

The analyses reported in this study were conducted using VACSR v.4.0 (https://cctvtt.com/vacsr4/) with the Stanza English POS model. The vocabulary list of the 402 words that appeared with a RANGE value of 2 in this study is available from the author upon request. The online version of VACSR retains uploaded texts and generated results only temporarily (approximately three hours) and automatically deletes them thereafter. An offline mode for local processing is planned for a future release.

References

Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-322.

Milton, J. (2009). *Measuring Second Language Vocabulary Acquisition*. Bristol, UK: Multilingual Matters.

Nation, I. S. P. (2001). *Learning Vocabulary in another Language*. Cambridge, UK: Cambridge University Press.

Nation, I. S. P. (2013). *Learning Vocabulary in another Language*. Cambridge, UK: Cambridge University Press.

Negishi, M., Takada, T., & Tono, Y. (2013). A progress report on the development of the CEFR-J. *The Language Teacher*, 37(4), 5-12.

Ohashi, Y., & Katagiri, N. (2020). The ratios of CEFR-J vocabulary usage compared with GSL and AWL in elementary EFL classrooms and suggestions of vocabulary items to be taught. *Asia Pacific Journal of Corpus Research, 1*(1), 61-94.

Ohashi, Y., Katagiri, N., & Honda, F. (2021). Classroom vocabulary analyzer combined with CEFR-J wordlist (CCVA): Tool development to examine vocabulary levels in classroom corpora based on the CEFR-J WORDLIST. *The International Journal of Language Learning & Applied Linguistics World, 27*(4) 1-12.

Ohashi, Y., & Katagiri, N. (2022). Vocabulary analyzer based on CEFR-J wordlist for self-reflection (VACSR): From classroom corpus compilation to self-reflection. *International Journal of Language Learning and Applied Linguistics World* (IJLLALW), *31*(1), 1-15.

Ohashi, Y., Katagiri, N., & Oshikiri, T. (2023). Vocabulary analyzer based on CEFR-J wordlist for self-reflection (VACSR) version 2. *Asia Pacific Journal of Corpus Research, 4*(2), 75-87.

Pintard, A., & François, T. (2020). Combining expert knowledge with frequency information to infer CEFR levels for words. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with Reading Difficulties* (READI 2020) (pp. 85-92). European Language Resources Association.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101-108). Association for Computational Linguistics.

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. New York: Palgrave Macmillan.

Takahashi, T. (2015). Vocabulary instruction and its impact on Japanese learners: Fostering deeper lexical knowledge and communicative competence. *Journal of Language Teaching and Research, 6*(3), 567-576.

Tono, Y. (Ed.). (2013). *The CEFR-J Handbook*. Tokyo: Taishukan Shoten.

THE AUTHOR

Yukiko Ohashi is a professor at Yamazaki University of Animal Health Technology. She earned her PhD in literature in 2014. Her principal research lies in corpus linguistics. She has published several articles on aspects of language learning, in particular corpus compilation.

THE AUTHOR'S ADDRESS

First and Corresponding Author
Yukiko Ohashi
Professor
Yamazaki University of Animal Health Technology
4-7-2 Minami-Osawa, Hachioji, Tokyo 192-0364, JAPAN
Email: y_watanabe@yamazaki.ac.jp

Received: 18 September 2025
Received in Revised Form: 9 December 2025
Accepted: 23 December 2025