Asia Pacific Journal of Corpus Research
Vol. 4, No. 2, 75-87

http://www.icr.or.kr/apjcr
e-ISSN 2733-8096
https://doi.org/10.22925/apjcr. 2023.4.2.75

Vocabulary Analyzer Based on CEFR-J Wordlist for Self-Reflection (VACSR) Version 2

Yukiko Ohashi
(Yamazaki University of Animal Health Technology)
Noriaki Katagiri
(Hokkaido University of Education)
Takao Oshikiri
(Bunkyo Gakuin University)

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.

This paper presents a revised version of the vocabulary analyzer for self-reflection (VACSR), called VACSR v.2.0. The initial version of the VACSR automatically analyzes the occurrences and the level of vocabulary items in the transcribed texts, indicating the frequency, the unused vocabulary items, and those not belonging to either scale. However, it overlooked words with multiple parts of speech due to their identical headword representations. It also needed to provide more explanatory result tables from different corpora. VACSR v.2.0 overcomes the limitations of its predecessor. First, unlike VACSR v.1, VACSR v.2.0 distinguishes words that are different parts of speech by syntactic parsing using Stanza, an open-source Python library. It enables the categorization of the same lexical items with multiple parts of speech. Second, VACSR v.2.0 overcomes the limited clarity of VACSR v.1 by providing precise result output tables. The updated software compares the occurrence of vocabulary items included in classroom corpora for each level of the Common European Framework of Reference–Japan (CEFR-J) wordlist. A pilot study utilizing VACSR v.2.0 showed that, after converting two English classes taught by a preservice English teacher into corpora, the headwords used mostly corresponded to CEFR-J level A1. In practice, VACSR v.2.0 will promote users' reflection on their vocabulary usage and can be applied to teacher training.

Keywords: Vocabulary, Classroom Corpus, CEFR-J, Teacher Training

1. Background

Since the introduction of using corpus linguistics as a means of analyzing patterns of language with different contexts, vocabulary-related research has produced several frequency-based vocabulary lists. Owing to the extensive array of corpus tools readily accessible through platforms such as Sketch Engine (Kilgarriff et al., 2014) and AntConc (Anthony, 2022), a robust foundation currently exists to pursue diverse strands of corpus-oriented research. Considering the close connection between vocabulary knowledge and overall language proficiency (Schmitt, 2010), structured vocabulary lists have been introduced to enhance vocabulary skills. Examples of such lists include the General Service List (GSL), pioneered by West (1958), and the Academic Word List (AWL), created by Coxhead (1998, 2000), aiming to promote effective vocabulary acquisition.

CEFR-J by Negishi, Takada, and Tono (2013), the Japanese version of the CEFR for Japanese students learning English, presented the CEFR-J wordlist with four levels: A1 (1,166-word types), A2 (1,411), B1 (2,445), and B2 (2,779). The CEFR-J can-do descriptors offer comprehensive information on the language skills and competencies learners are expected to possess at each CEFR-J proficiency level. The CEFR-J word list assists teachers in effectively incorporating these can-do descriptors into their

teaching, helping them select appropriate vocabulary to align with the standards outlined in the Education Ministry guidelines. The CEFR-J wordlist is the benchmark of vocabulary items that language teachers use or teach in a classroom. Corpus-based wordlists based on abundant corpus evidence enable curricula design and content-based tasks that allow English learners to manage unknown vocabulary items while reading.

Employing a corpus-supported vocabulary inventory enables the creation of indispensable lexical compendiums tailored for diverse domains. As an illustrative instance, the scholarly investigation conducted by Ohashi and Katagiri (2020) scrutinized the proportions of CEFR-J vocabulary utilization compared with the General Service List (GSL) and the Academic Word List (AWL) within primary-level English as a Foreign Language (EFL) classrooms. The study suggested that educators at the elementary level might be inclined to restrict their students' exposure to syntactically refined constructs that are abundant in lexical diversity. This tendency stems from educators' recurrent utilization of a restricted lexicon during instructional sessions.

Vocabulary research endeavors frequently demand corpora construction. The assembly of classroom corpora entails a meticulous manual tagging procedure susceptible to human errors. In response to this challenge, Ohashi, Katagiri, and Oshikiri (2022a) introduced the Classroom Corpus Tagger (CCT). This innovative tool was devised to autonomously allocate speaker and language tags to discrete statements, with the overarching objective of alleviating the logistical or practical quandaries associated with the compilation of classroom corpora. The tool aimed to facilitate educators' reflections to enhance their pedagogical practices.

1.1. Developing and Upgrading

Combining classroom corpora with an analysis involving a corpus-based vocabulary list allows language teachers to reflect on their vocabulary usage because it reveals the levels and occurrences of the vocabulary items they use. The vocabulary analyzer, based on the CEFR-J wordlist for self-reflection (VACSR), which complies with the CEFR-J wordlist developed by Ohashi, Katagiri, and Oshikiri (2022b), aims to reveal which vocabulary items are covered in classrooms at each level to facilitate teachers' reflection on vocabulary teaching. VACSR v.1.0 enables the comparison of multiple classroom corpora. This tool can compare the number of word occurrences between different files for vocabulary items included in the CER-J wordlist. The analysis results are displayed in separate columns for each file, making it easy to compare the number of word occurrences across multiple files. However, VASCR v.1.0 had limitations to address. First, it does not distinguish 1) words with identical spelling that are different parts of speech (POS). For example, the word *mean* in the CEFR-J wordlist has three headwords in as many POS and CEFR-J levels; *mean* (verb, A1), *mean* (adjective, A2), and *mean* (noun, B1). VACSR v.1.0 would tabulate a single occurrence of "mean" in one count in *mean* (A1), *mean* (A2), and *mean* (B1), causing confusion. Second, VASCR v.1.0 does not identify multiple-word tokens (MWTs) such as "bad-tempered," "brother-in-law," and "well-known," whereas the CEFR-J wordlist specifies such 62 MWTs as distinctive lexical items in the wordlist. Leaving out the 62 MWTs would more or less affect the analysis results and user's interpretation. These limitations result in ambiguous outputs. The authors updated VACSR v.1 to VACSER v.2.0 by equipping it with the capabilities to overcome the two limitations. In this paper, the authors present the updated VACSR v.2.0 from v.1.0 and showcase the pilot research findings obtained through its application.

2. Method

2.1. Justification for the Selection of Stanza for POS Parsing

When selecting an appropriate tool for POS tagging in VACSR v.2.0, our primary consideration was

accuracy, leading us to choose Stanza, developed by the Stanford NLP Group. Stanza's models stand at the forefront of POS tagging performance. Leveraging advanced deep learning techniques, they have been meticulously trained on extensive datasets, resulting in consistently high accuracy rates. Stanza boasts pre-trained neural models that support 70 human languages, ensuring that VACSR v.2.0 is built on a foundation of reliable data.

Furthermore, Stanza is freely available for use, setting it apart from proprietary POS tagging tools in the market, such as Amazon Comprehend, Google Natural Language AI, and Microsoft Cloudmersive NLP, which are commercial products.

2.2. How to Use VACSR V.2.0

This section describes how to use VACSR v.2.0 and what it can produce. Users input *.txt* or *.xml* files into VACSR v.2.0 for analysis, and it can generate output files in a *.csv* format (Figure 1).

Figure 1. The Process of VACSR 2 Producing Its Output

VACSR v.2.0 can read several corpora simultaneously and compare word occurrences and vocabulary levels between different files according to the CEFR-J wordlist. Figure 2 shows the initial viewing screen. The users follow three steps to obtain the analysis results. VACSR v.2.0 is available at https://cctvtt.com/vacsr2/.

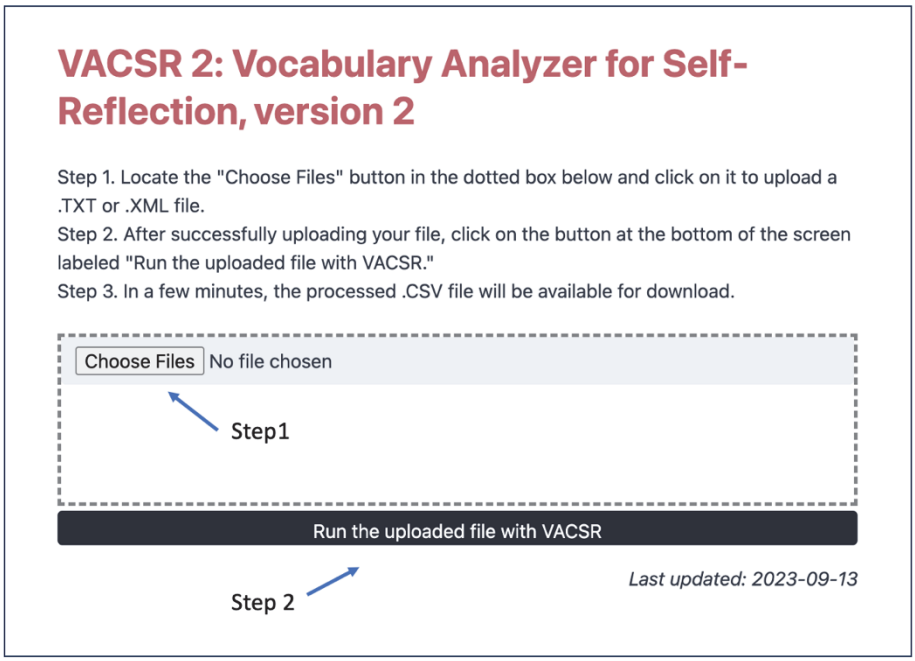

Figure 2. VACSR 2 Viewing Screen

Step 1: Upload .txt or .xml files from “Choose Files.”
Step 2: Press the “Run the uploaded file with VACSR” button to execute the uploaded file.
Step 3: A few minutes after execution, press the refresh button on the browser and generate a result file in the specified directory.

There are two options for downloading the result output files: an output file or one containing multiple POS. These two options are accessed by clicking “Download Processed File” and “Download Multiple Parts of Speech.” These files are inaccessible three hours after generating the result files.

3. Update to VACSR V.2.0

One of the limitations that VACSR v.1.0 faced was headwords with multiple POSs, and the other was MWTs, whether they were hyphenated or not. We revised VACSR v.1.0 to v.2.0, overcoming those limitations: dissecting multiple parts of speech from a single headword and identifying word MWTs, which were hyphenated or non-hyphenated. The following subsections explain how we overcame these limitations and achieved updates.

3.1. Multiple Parts of Speech with a Single Headword

To avoid multiple occurrences of a headword with multiple POS, VACSR v.2.0 used Stanza (Qi et al., 2020) to parse the file containing target passages contains target passages, namely, to parse POS for preconditioning sentences in the target file to be analyzed.

Stanza’s POS list contains 36 POS possibilities, while the CEFR-J wordlist uses 15 POSs, 21 fewer than Stanza’s. Some POS tags are identical or very similar, such as “determiner” and “adjective.” Other POS tags, such as Stanza’s “personal pronouns” and “possessive pronouns,” convertible to “pronouns” are easily converted. However, others have more intricate coverage of POS parsing than CEFR-J. We identified 29 Stanza POSs as ten corresponding CERF-J POS equivalents. We did not identify the CEFR-J's POS equivalents in the remaining 7 Stanza’s POSs, such as “FW, foreign word,” “LS, list item marker,” and “NNP, proper noun singular.” We disregarded such POSs because the CEFR-J wordlist does not categorize them. Table 1 shows a list of POS conversions from Stanza to CEFR-J.

Table 1. Conversion Table of Stanza and CEF-J POS Equivalents

POS Number	POS Tag Name	Description in Stanza	➔	CEFR-J POS Equivalent Tag Name
1	CC	Coordinating conjunction		conjunction
2	CD	Cardinal number		number
3	DT	Determiner		determiner
4	EX	Existential there		adverb
5	FW	Foreign word[a]		N/A
6	IN	Preposition or subordinating conjunction		Preposition
7	JJ	Adjective		adjective
8	JJR	Adjective, comparative		adjective
9	JJS	Adjective, superlative		adjective
10	LS	List item marker[a]		N/A
11	MD	Modal		modal auxiliary
12	NN	Noun, singular, or mass		noun
13	NNS	Noun, plural		noun
14	NNP	Proper noun, singular[a]		N/A
15	NNPS	Proper noun, plural[a]		N/A
16	PDT	Predeterminer		determiner
17	POS	Possessive ending[a]		N/A
18	PRP	Personal pronoun		pronoun
19	PRP$	Possessive pronoun		pronoun

20	RB	Adverb	adverb
21	RBR	Adverb, comparative	adverb
22	RBS	Adverb, superlative	adverb
23	RP	Particle[b]	N/A
24	SYM	Symbol[a]	N/A
25	TO	to	infinitive-to
26	UH	Interjection	interjection
27	VB	Verb, base form	verb
28	VBD	Verb, past tense	verb
29	VBG	Verb, gerund, or present participle	verb
30	VBN	Verb, past participle	verb
31	VBP	Verb, non-third person singular present	be-verb, verb
32	VBZ	Verb, third person singular present	be-verb, verb
33	WDT	*Wh*-determiner	determiner
34	WP	*Wh*-pronoun	pronoun
35	WP$	Possessive *wh*-pronoun	adverb or determiner[c]
36	WRB	*Wh*-adverb	adverb

Note. POS = part of speech. CEFR-J = Common European Frame of Reference–Japanese. N/A = not applicable.
[a] The CEFR-J wordlist does not include these categories. [b] Particles such as *up* and *out* used in "look up, and "check out" are parsed as RP (particle) in Stanza, while the CEFR-J wordlist does not provide the RP as its POS category. They are parsed as "adverb" in VACSR v.2.0. [c] Possessive *wh*-pronoun *whose* in the CEFR-J wordlist is classified as either an adverb or determiner. They were both CEFR-J level A1. Thus, they do not affect the VACSR v.2.0 result.

3.2. Multiple Word Tokens (MWTs)

Stanza is not equipped with MWTs in English. Some MWTs are hyphenated, for example, *bad-tempered* and *brand-new*, whereas others are not, *according to* and *because of*. Thus, we first listed the hyphenated words that the CEFR-J wordlist designates and preset their CEFR-J levels (Table 2). The CEFR-J wordlist contains 62 hyphenated MWTs. Fifty-seven hyphenated MWTs out of 62 MWTs possess a singular POS and are thus directly annotated with their respective POSs. The remaining five hyphenated MWTs exhibited multiple parts of speech.

Table 2. Samples of Hyphenated Headwords in the CEFR-J Wordlist

Headword Sample*	POS	CEFR-J Wordlist Level
bad-tempered	adjective	B2
brand-new	adjective	B1
brother-in-law	noun	B2
CD-ROM	noun	B1
check-in counter/check-in	noun	B1
check-in desk/check-in	noun	B1
daughter-in-law	noun	B2
duty-free	adjective	B1

Note. POS = parts of speech. CEFR-J = Common European Framework of Reference for Languages–Japanese. * The CEFR-J wordlist contains 62 hyphenated words.

We identified six non-hyphenated MWTs in the CEFR-J wordlist. They were prepositional phrases—*according to, because of, instead of, next to, on to,* and *owing to*—were categorized as "prepositions." Thus, we recategorized these prepositional phrases as "prepositions" before parsing with Stanza.

3.3. VACSR V.2.0 Processing Flow

The preceding section explained how to manage the two limitations of the VACSR predecessor: (1) headwords with different POSs and (2) MWTs. We designed preprocessing steps to extract hyphenated MTWs and prepositional phrases from the input sentences. Preprocessing creates a wordlist that contains four CEFR-J levels ranging from A1 (*primary*) to B2 (*advanced*) before Stanza's annotation. After preprocessing, VACSR v.2.0 parses the remaining sentences using Stanza. Figure 3 shows the VACSR v2.0 input-processing-output (IPO) cycle.

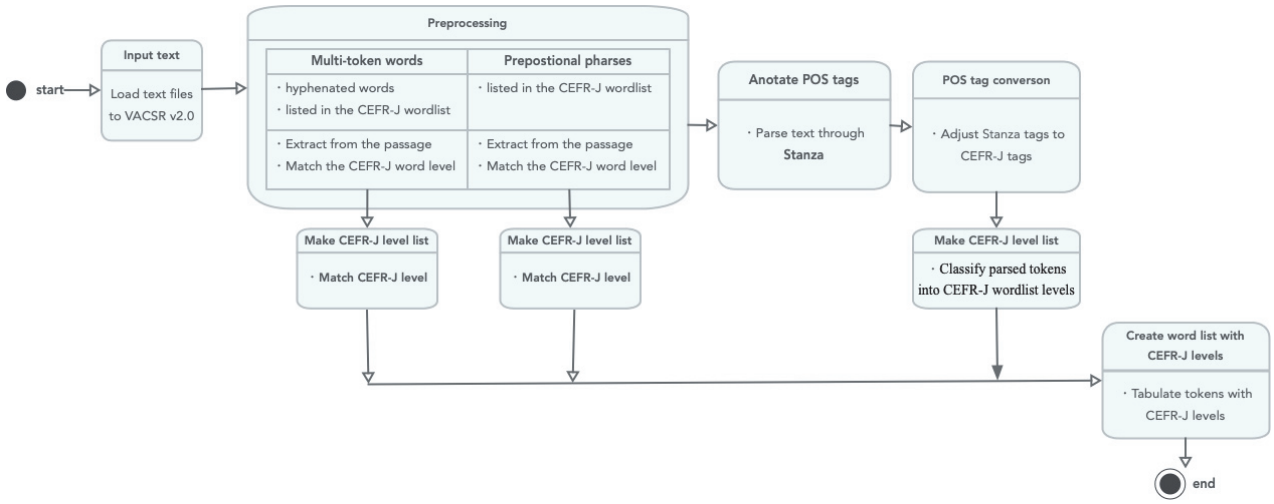

Figure 3. IPO Cycle of VACSR v.2.0 with Preprocessing and Stanza Parsing

The updated VACSR v.2.0 enables categorizing lexical items with multiple POS, providing the analyzed results. For example, the word *mean* can be a verb, an adjective, or a noun. Figure 4 displays an excerpt of the output. The five MWTs (*full-time*, *grown-up*, *half-price*, *part-time*, and *second-hand*) with multiple POS require manual classification. A separate "multi-pos" file is outputted to show the occurrences of *grown-up* and *part-time*. If they are not present, an empty file is generated.

Verb categorized in level A1

Headword	POS	CEFR-J	RANGE	FREQ	Corpus 1	Corpus 2
matter	noun	A1	1	1	0	1
me	pronoun	A1	1	3	3	0
mean	verb	A1	1	1	0	1
message	noun	A1	1	1	0	1

Adjective categorized in level A2

Headword	POS	CEFR-J	RANGE	FREQ	Corpus 1	Corpus 2
marry	verb	A2	1	1	1	0
mean	adjective	A2	1	1	0	1
next	adverb	A2	2	6	4	2
part	noun	A2	1	1	0	1

Noun categorized in level B2

Headword	POS	CEFR-J	RANGE	FREQ	Corpus 1	Corpus 2
mean	noun	(B2)	0	0	0	0
mechanize /mechanise	verb	(B2)	0	0	0	0
media	noun	(B2)	0	0	0	0

"mult-pos" results

Corpus	Headword	Concordance
1.txt	grown-up	of transportation. she is a grown-up adult. I have determined
1.txt	part-time	have determined to quit a part-time job. she took a train.

Figure 4. VACSR 2.0 Results of 'Mean' with Multiple POS

3.4. Clarity Issue of Showing Uncovered Vocabulary Items

VACSR v.1.0 displayed the uncovered vocabulary items at the end of the list, making it cumbersome to compare covered and uncovered words for each level (A1, A2, B1, and B2). In contrast, VACSR v.2.0 showcases uncovered vocabulary items level-wise in parentheses as (A1), (A2), (B1), or (B2). This alteration is one of the modifications implemented in VACSR v. 2.0. Table 3 lists the samples. The parenthetically labeled words *A.M., about, above, action*, and *activity* were not used in the classroom by teachers or students in this example. This function enables teachers to notice tendencies in their vocabulary usage and provides a source for teachers' self-reflection on their vocabulary use in the classroom.

Table 3. VACSR 2.0 Non-Existent Token Display Samples at CEFR-J level in Parentheses

Headword	POS	CEFR-J	RANGE	FREQ	Corpus 1
word	noun	A1	1	1	1
work	noun	A1	1	1	1
you	pronoun	A1	1	3	3
your	determiner	A1	1	2	2
yourself	pronoun	A1	1	1	1
a.m./A.M. /am/AM	adverb	(A1)	0	0	0
about	preposition	(A1)	0	0	0
above	preposition	(A1)	0	0	0
action	noun	(A1)	0	0	0
activity	noun	(A1)	0	0	0

Note. POS = parts of speech. CEFR-J = Common European Framework of Reference for Languages–Japanese. FREQ = frequency.

3.5. Vocabulary Items Shown as (other) in the CEFR-J Wordlist

All vocabulary items applied to the four cases below are shown as (other) below (B2). Table 4 presents the sample results.

1) Plural forms of the nouns (e.g., dishes, textbooks, reports)
2) The third person presents singular verbs (e.g., washes, takes, means)
3) Past forms, passive forms (e.g., took, talked, protected)
4) All the other words that are not included in the CEFR-J wordlist (e.g., “last” as an adjective)

Table 4. Determiners in CEFR-J Wordlist

Headword	POS	CEFR-J	RANGE	FREQ	Corpus 1
yummy	adjective	(B2)	0	0	0
zebra	noun	(B2)	0	0	0
zip	noun	(B2)	0	0	0
zoom	noun	(B2)	0	0	0
brought	verb	(other)	1	1	1
buses	noun	(other)	1	1	1
discusses	verb	(other)	1	1	1
dishes	noun	(other)	1	1	1
disliked	verb	(other)	1	1	1
items	noun	(other)	1	1	1
last	adjective	(other)	1	1	1
listed	verb	(other)	1	1	1
means	verb	(other)	1	1	1
next	adjective	(other)	1	2	2
points	verb	(other)	1	1	1
said	verb	(other)	1	1	1
takes	verb	(other)	1	1	1
talks	verb	(other)	1	1	1
tasks	noun	(other)	1	1	1

Note. POS = parts of speech. CEFR-J = Common European Framework of Reference for Languages–Japanese. FREQ = frequency.

4. Pilot Study Using VACSR V2.0 and Results

The preceding sections explained the updates to VACSR from v.1.0 to v.2.0 by addressing two issues: (1) multiple parts of speech with a single headword and (2) multiple word tokens (MWTs). This section presents a pilot study using VACSR v.2.0, utilizing the output results of VACSR v.2.0. The objective of the pilot study is to illustrate how VACSR v.2.0 can be utilized by prospective VACSR users. Thus, the present pilot study did not intend to draw any conclusive finding or remark in the corpus linguistics due to the limited small dataset size of the corpora that the study used but aimed to display how VACSR v.2.0 would yield vocabulary analysis results as the authors had intended its update from VACSR v.1.0.

4.1. Participant and Dataset

The material subcorpora were created from one preservice English teacher’s teaching practice English lessons at an elementary school and at a junior high school. The classes were recorded and transcribed. Only the preservice teacher’s English utterances were extracted and converted into text files. The authors prepared the dataset for the sake of using it on VACSR v.2.0 to extract more precise vocabulary analysis results with CEFR-J wordlist POS information parsed by Stanza.

4.2. Aims

VACSR v.2.0 distinguishes the word levels in the text file input from multiple corpora based on the CEFR-J wordlist's four levels: A1 (*elementary*) to B2 (*advanced*). Non-appearing words are displayed in parentheses, although they exist in the CERF-J wordlist. As a starter, the pilot study examined the results from VACSR v.2.0, which can be compared and used for analyses and reflections regarding CEFR-J word levels. We posed the following research questions (RQs):

RQ 1. Do the preservice teacher's word levels differ depending on school types?
RQ 2. Does the preservice teacher exhibit common vocabulary items between the two types of schools?

4.3. Results and Discussion

VACSR v.2.0 analyzed two text files, one from an elementary school English class and the other from a junior high school English class. The pilot study summarized CEFR-J level occurrences among these two school types depending on the ranges. In the pilot study, "Range 2" lists headwords that appeared in both corpora and "Range 1" lists headwords that appeared in either of the two corpora because the pilot study used two corpora. Tables 5 and 6 show the occurrences of CEFR-J wordlist headword types in levels A and B, respectively.

Table 5. CEFR-J Wordlist Level A Headword Type Occurrences by a Preservice English Teacher

Corpus	CEFR-J Wordlist Level			
	A1		A2	
	Range		Range	
	2	**1**	**2**	**1**
	Headword Type Occurrences			
Elementary School	49	47	6	2
Junior High School		116		8

Table 6. CEFR-J Wordlist Level B Headword Type Occurrences by a Preservice English Teacher

Corpus	CEFR-J Wordlist Level			
	A1		A2	
	Range		Range	
	2	**1**	**2**	**1**
	Headword Type Occurrences			
Elementary School	1	5	3	2
Junior High School		13		7

Regarding the CEFR-J wordlist levels, A1 displayed more occurrences than the other levels. A2, B1, and B2 show fewer occurrences than A1 in elementary and junior high schools. Examining the occurrences in range 1 in all four levels, the junior high school displayed more occurrences than elementary school, especially in A1. This suggests that the preservice teacher adjusted the word variety according to school type. Focusing on the occurrences in range 2 in A1, forty-nine was the largest among all ranges (Tables 5 and 6). Table 7 presents a list of the top 20 CEFR-J word level A1 headwords in Range 2, i.e., the 20 words in Table 7 appeared in both of the corpora.

Table 7. Top 20 Headwords Appearing in Range 2 in CEFR-J Level 1

Headword	POS	CEFR-J Level	RANGE	FREQ	Corpus 1 Freq Elementary School	Corpus 2 Freq Junior High School
you	pronoun	A1	2	134	66	68
I	pronoun	A1	2	101	54	47
is	be-verb	A1	2	89	37	52
your	determiner	A1	2	63	14	49
like	verb	A1	2	55	52	3
it	pronoun	A1	2	53	16	37
the	determiner	A1	2	48	7	41
and	conjunction	A1	2	41	9	32
one	determiner	A1	2	36	3	33
thank	verb	A1	2	34	33	1
do	do-verb	A1	2	30	18	12
two	number	A1	2	28	3	25
a	determiner	A1	2	28	4	24
three	number	A1	2	24	3	21
yes	adverb	A1	2	24	3	21
we	pronoun	A1	2	23	12	11
ready	adjective	A1	2	21	6	15
have	have-verb	A1	2	20	3	17
what	determiner	A1	2	19	18	1
September	noun	A1	2	18	8	10

Note. POS = parts of speech. CEFR-J = Common European Framework of Reference for Languages–Japanese. FREQ = frequency.

Considering that most of these words represent function words, basic words, and A1 to be taught at an elementary level (Table 7), we can assume that the preservice teacher used them as working vocabulary items regardless of the school type. The higher word levels than A1 did not attract as many words.

Judging from the results of VACSR v. 2.0, we obtained evidence to provide answers to the RQs. On the one hand, the skewed headwords in Range 2 in A1 seem to indicate that core vocabulary pertains to teaching in both elementary and junior high schools. It answered RQ2. On the other hand, considering the occurrences in Range 1 at all levels, the preservice teacher's usage differed depending on the school types (Tables 5 and 6). The preservice teacher's word usage varied more widely when teaching a junior high school English class than when teaching an elementary school class. This answers RQ1.

Examining all the results, for example, the distributions of the headwords and unused words, shown in parentheses in the results, the preservice teacher can reflect on the word usage and be assisted in improving and expanding the word coverage despite the CEFR-J word levels. We must remember in drawing a conclusion that the sizes of the corpora were very limited due to the nature of the present pilot study. The pilot study used a limited amount of text to analyze with VACSR v.2.0. More voluminous text files would have provided more insightful results. We must rely on further analyses to confirm the pilot study findings with larger size corpora compiled from more English classes that will be recorded for a longer period of time with more teachers and different schools involved. However, this section has demonstrated one possible application of utilizing the results obtained from VACSR v. 2.0.

5. Implications

With additional functions that distinguish headwords with different POSs, VACSR v.2.0 can identify POS, producing result output tables that are relatively more precise than those in the literature. VACSR v.2.0 can also classify words with identical spellings with different POSs. Using VACSR v.2.0 for CEFR-J vocabulary level analyses benefits its users.

VACSR v.2.0 output tables provide users with occurrences and non-occurrences of vocabulary items in the CEFR-J wordlist accompanied by their respective word levels (Tables 3 and 4). This helps VACSR v.2.0 users reflect on their vocabulary usage. Combined with multiple corpora, VACSR v.2.0 facilitates a comparison of vocabulary usage according to each level in the CEFR-J wordlist, which provides materials for self-reflection. This will contribute to improving vocabulary usage in classrooms.

The pilot study illustrated a case of vocabulary-level analysis of a preservice teacher's English classes. However, depending on the users' contexts, VACSR v.2.0 can be applied to other cases: students learning English, in-service English teachers, English teacher educators, and researchers.

We created VACSR v.2.0 because we needed to comprehend word occurrences in English texts according to the CEFR-J vocabulary levels. The prospective utility of VACSR in forthcoming updates will equip the system with different languages that CEFR refers to, for example, French and Spanish. If we obtain their CEFR vocabulary lists, such as the CEFR-J wordlist, its functionality extends to quantifying and contrasting the occurrence of vocabulary elements from diverse languages. This tool is specifically designed to analyze English vocabulary items; therefore, linguistic components not existing in English are subsequently categorized as "other," in alignment with the CEFR-J level B2.

An illustrative example of this process is shown in Figure 5. If the spelling of words perfectly aligns with the English language, they are categorized according to their respective proficiency levels. This will expand the prospective utility of VACSR to accommodate diverse languages other than English in the future.

Words categorized in level A2

Headword	POS	CEFR-J	RANGE	FREQ	Corpus 1	Corpus 2
intelligent	A2	2	2	10	187102	52.21714
international	A2	2	2	117	1527432	610.94054
local	A2	2	2	50	1208695	261.0857

Words categorized in level B2

Headword	POS	CEFR-J	RANGE	FREQ	Corpus 1	Corpus 2
indispensable	B2	2	2	15	410329	78.32571
initial	B2	2	2	7	270276	36.551998
invisible	B2	2	2	6	102670	31.330284

Words not encompassed within any specific group

Headword	POS	CEFR-J	RANGE	FREQ	Corpus 1	Corpus 2
vulnérable	(other)	2	2	12	67630	62.660568
véritable	(other)	2	2	40	812637	208.86856
âgé	(other)	2	2	21	371847	109.65599
écologique	(other)	2	2	23	247872	120.09942
économique	(other)	2	2	77	1315685	402.07198
complémentaire	(other)	2	2	7	276614	36.551998
compétent	(other)	2	2	7	151150	36.551998

Figure 5. VACSR 2.0 Results of French Words

References

Anthony, L. (2022). AntConc (Version 4.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Coxhead, A. (1998). *The development and evaluation of an academic word list* (Master Thesis, Victoria University of Wellington, New Zealand).

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The sketch engine: Ten years on. *Lexicography*, *1*, 7-36.

Negishi, M.,Takada, T., & Tono, Y. (2013). A progress report on the development of the CEFR-J. In E.D. Galaczi & C. J. Weir (Eds.), Exploring language frameworks. *Proceedings of the ALTE Krakow Conference*, July 2001, 135-163.

Ohashi, Y., & Katagiri, N. (2020a). 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), 35-65.

Ohashi, Y., Katagiri, N., & Oshikiri, T. (2022b). Developing classroom corpus tagger for teachers' reflective practice: A spoken language tagger to compile classroom corpora. *English Corpus Studies*, *29*, 41-62.

Ohashi,Y., Katagiri, N., & Oshikiri, T. (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, 31*(1), 1-15.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. (2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

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

West, M. (1953). *A General Service List of English Words*. Longman, London.

Penn Treebank P.O.S. Tags. (n.d.). www.ling.upenn.edu. https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

THE AUTHORS

Yukiko Ohashi is an Associate Professor at Yamazaki University of Animal Health Technology. Her principal research lies in the field of corpus linguistics.

Noriaki Katagiri is a Professor at Hokkaido University of Education. His research interests include spoken corpora, classroom discourse analyses, and English language acquisition.

Takao Oshikiri is an Associate Professor at Bunkyo Gakuin University. In 2003, he earned his MSc in International Business from London South Bank University. He has authored books in the field of digital marketing.

THE AUTHOR'S ADDRESS

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

Co-author
Noriaki Katagiri
Professor

Hokkaido University of Education, Asahikawa
9 Chome, Hokumoncho, Asahikawa, Hokkaido 070-8621, JAPAN
Email: katagiri.noriaki@a.hokkyodai.ac.jp

Co-author
Takao Oshikiri
Associate Professor
Bunkyo Gakuin University
1-19-1 Mukogaoka, Bunkyo, Tokyo 113-8668, JAPAN
E-mail: toshikiri@bgu.ac.jp

Received: 1 October 2023
Received in Revised Form: 17 November 2023
Accepted: 10 December 2023