CLASSROOM VOCABULARY ANALYZER COMBINED WITH CEFR-J WORDLIST (CCVA): TOOL DEVELOPMENT TO EXAMINE VOCABULARY LEVELS IN CLASSROOM CORPORA BASED ON THE CEFR-J WORDLIST

Yukiko Ohashi
Yamazaki University of Animal Health Technology
Fuhito Honda
Graduate of The University of Tokyo
Noriaki Katagiri
Hokkaido University of Education

ABSTRACT
The current study introduces a vocabulary analysis tool we developed based on the CEFR-J wordlist in order to identify variations in the levels of vocabulary items in classroom corpora. The tool, named Classroom Corpus Vocabulary Analyzer with CEFR-J Wordlist (CCVA), allows the vocabulary levels used in a given classroom to be analyzed, revealing overused vocabulary items and words that are less likely to appear in classroom settings, with the aim of suggesting a wider range of vocabulary usage in the classroom. The study shows the procedure of developing CCVA, including its programming design, followed by a pilot study that discussed the availability of CCVA. The pilot study revealed that the CCVA precisely extracted the target classroom utterances and interpreted the annotated classroom corpora in XML format, producing the English tokens’ distribution on the CEFR-J’s four scales (A1, A2, B1, and B2). In the pilot study, the teachers’ use of the English language revealed that the CEFR-J level A1 accounted for the most significant ratios. A sharp drop occurred in Level A2, and the drop changed to a gradual decrease in B1 and B2. This study presents how we use CCVA and discusses how we interpret the gained outcomes with the tool.

KEYWORDS
Corpus Linguistics, CEFR-J, vocabulary, teacher education, classroom corpora

INTRODUCTION
Since the publication of The Common European Framework of Reference for Languages (CEFR) in 2001, the CEFR has become an influential standard for designing language curricula,
providing a guideline for learning, teaching, and assessing learners' ability. From the onset of CEFR, it has acted as a practical descriptive scheme that can be used for analyzing the needs, goals, materials, and outcomes of language learners’ studies (Alanen, Huhta & Tarnanen, 2010).

The CEFR describes itself as being a comprehensive basis for establishing language syllabuses and curriculum guidelines for the elaboration of teaching and learning materials (Council of Europe, 2014). As CEFR frameworks have influenced countries beyond Europe to East Asia, applying their frameworks in Asian educational settings, where English is still a foreign language (EFL), has revealed some conceptual problems. Wu (2012) presented an adjusted version of the CEFR based on the General English Proficiency Test (GEPT) to fit the Taiwanese context, which highlighted the difficulties in assessing learners’ abilities based on CEFR frameworks.

Read (2019) argues that English language educators must understand the underlying principles of CEFR and adjust them in their teaching in ways that are socially and culturally appropriate according to the context.

To face the challenges in applying European CEFR frameworks to Japanese students, CEFR-J (Common European Framework of Reference-Japan, Negishi, Takada & Tono, 2013) was introduced in Japan. The CEFR-J project started in 2008, compiling the original framework based on the CEFR for English language teaching in Japan (Negishi, Takada, and Tono 2013; Tono 2013; Negishi and Tono 2016). The establishment of CEFR-J led to the presentation of a new vocabulary list that could be used as a measure to analyze vocabulary level. This study attempts to develop a vocabulary level analysis tool based on the CEFR-J combined with classroom corpora.

**LITERATURE REVIEW**

**CEFR-J**

The CEFR-J is a modified version of the CEFR and was designed as an original framework specific to the context of teaching English in Japan. While the CEFR has a six-level scale (A1, A2, B1, B2, C1, and C2), CEFR-J has sub-levels of the CEFR (Pre-A1, A1.1-1.3, A2.1-2.2, B1.1-1.2, and B2.1-2.2). The CEFR-J provides newly designed descriptors, describing the communicative competencies of a learner by using five language sub-skills (listening, reading, spoken interaction, spoken production, and writing). The five skills the CEFR-J CAN DO lists (listening, reading, spoken interaction, spoken production, and writing) accompanied by the CEFR-J wordlist are designed to accommodate Japanese English learners, most of whom belong to the A1 level (Negishi, et al., 2013).
Following the introduction of the CEFR-J, expanded frameworks have been established for other countries as well, for example, the CSE: China’s Standards of English Language Ability (Tono & Negishi, 2020). Given that other language tests (e.g., Cambridge English Qualifications, International English Language Testing System) have linked their standards to the CEFR, additional CEFR standards exclusive to different countries could be developed in the same way as the CEFR-J.

**CEFR-J wordlist**

The establishment of the CEFR-J also resulted in the development of the CEFR-J wordlist. The current version, ver. 1.6, is available at http://www.cefr-j.org/download.html#cefrj_wordlist. The descriptions of the CEFR-J wordlist are also presented on the website of a group of researchers from Tono Laboratory at Tokyo University of Foreign Studies (TUFS). The CEFR-J wordlist was developed based on the English textbook corpora comprising the major English textbooks used at primary to secondary schools (Years 3 to 10) in China, Korea, and Taiwan (Tono, 2013). Table 1 shows the total vocabulary items pertaining to each level of the CEFR-J wordlist. Some words are repeated at different levels depending on their parts of speech (POS). For example, referring to the word “above,” it is included in level A1 as an adjective and as an adverb in level B1. The CEFR-J wordlist facilitated related research, including specific attempts to create online applications. For example, Uchida and Negishi (2019) presented CVLA 2 (CEFR-based Vocabulary Level Analyzer), which assigns CEFR-J levels (Pre-A1 to C2) to English text based on textual features calculated from the input text. CVLA enables to examine textual and grammatical features based on the CEFR-J standard.

**Table 1: Total Numbers of Vocabulary Items on Each Level in CEFR-J Wordlist**

<table>
<thead>
<tr>
<th>Word level</th>
<th>Total number</th>
<th>Descriptions of vocabulary levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>1,164</td>
<td>From beginner to lower-intermediate</td>
</tr>
<tr>
<td>A2</td>
<td>1,411</td>
<td>From upper-beginner to lower-intermediate</td>
</tr>
<tr>
<td>B1</td>
<td>2,446</td>
<td>From lower-intermediate to intermediate</td>
</tr>
<tr>
<td>B2</td>
<td>2,778</td>
<td>From upper-intermediate to advanced</td>
</tr>
<tr>
<td>All levels</td>
<td>7,799</td>
<td></td>
</tr>
</tbody>
</table>

Although the CEFR-J has become widely known, Vocabulary analysis research based on the CEFR-J wordlist is still developing. Referring to a recent study, Ohashi and Katagiri (2020) presented broad ratios of vocabulary levels observed in classroom corpora based on the CEFR-J.
The implications are that the usage of vocabulary items belonging to CEFR-J level A1 is rare in Japanese elementary schools. Due to the complexity of the analyses and counting the ratios using statistical tools, vocabulary analysis based on the CEFR-J wordlist combined with classroom corpora is comparatively challenging.

Applying the CEFR-J benchmark to classroom corpora, we can gain suggestions as to which level of vocabulary items are most likely to appear in each classroom or which level of vocabulary items should be introduced in a classroom. With the attempt to simplify CEFR-J-based vocabulary analysis, this study developed a Classroom Corpus Vocabulary Analyzer with the CEFR-J Wordlist (CCVA).

**METHODOLOGY**

*Tool Design for analyzing vocabulary level based on CEFR-J wordlist*

We created the web application for the vocabulary analyzer (figure 1) using the package, *Shiny*, included in the statistical processing software *R*. The newly created classroom corpus vocabulary analyzer is available at https://fuhitohonda.shinyapps.io/cefrj_analysis/.

First, we must insert tags attached to the utterances we hope to analyze in ① (Figure 1) followed by selecting the classroom corpus with annotations, which contains the utterances to be analyzed in ② (Figure 1). Then, raw data without annotation (sample data shown in Figure 2, the data that appear after clicking A), the CEFR-J vocabulary levels ratios for the words included in each utterance (sample shown in Figure 3, the data shown after clicking B ), and the graph of ratios (Figure 4, the data shown after clicking C) will appear. Figure 5, which appears after clicking D, shows the CEFR-J vocabulary usage ratios for the class from the beginning to the end. The analysis results can also be downloaded. Figure 6 is the annotated classroom corpus sample. Classroom corpora should be compiled in XML format. For example, suppose you want to extract the teacher’s English utterances to analyze them through the CEFR-J vocabulary level. In that case, you should insert “teacher/English” in the tool (see Figure 1) and select the classroom corpus you want to analyze. Then, vocabulary analysis results will show the variation of English vocabulary levels of the words the teacher used in a classroom.
Yukiko Ohashi, et al

Figure I: Web Application of Vocabulary Analyzer

CEFR-J Analysis

Figure II: Raw Data Shown in the Application

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>元データ</td>
<td>辞書適用後のデータ</td>
<td>全体での割合</td>
<td>レベル別出現割合の推移</td>
</tr>
<tr>
<td>value</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Good, morning, miss.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>How are you today?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>last time, we read ET story, so first, let’s try shadowing, ok, let’s start</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ok, Maho, what did ET say?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ok, he said, ‘I’ll be right here.’ ok, good</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ok, everybody, this time, open your textbook, page 75. Let’s try ‘read and look.’ sentence tc go.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>one more time, this sentence, go.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Yukiko Ohashi, et al
www.ijllalw.org

Figure III: Ratios of CEFR-J Vocabulary Levels

<table>
<thead>
<tr>
<th>doc_id</th>
<th>Types</th>
<th>Tokens</th>
<th>Sentences</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>Text</th>
<th>A1 proportion</th>
<th>A2 proportion</th>
<th>B1 proportion</th>
<th>B2 proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>6</td>
<td>8</td>
<td>2</td>
<td>3.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>Good morning, miss. oo</td>
<td>37.50</td>
<td>0.00</td>
<td>0.00</td>
<td>12.5%</td>
</tr>
<tr>
<td>2.00</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>4.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>How are you today?</td>
<td>80.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3.00</td>
<td>15</td>
<td>19</td>
<td>1</td>
<td>3.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>last time, we read ET story so first, let's try shadowing, ok, let's start!</td>
<td>47.37</td>
<td>15.79</td>
<td>6.26</td>
<td>5.26</td>
</tr>
</tbody>
</table>

Figure IV: Graph of CEFR-J Vocabulary Ratios

Figure V: CEFR-J Vocabulary Ratios Observed in a Class

The international journal of language teaching and applied linguistics world
www.ijllalw.org
Copyright 2021
How to Analyze

R's text metric analysis package *quanteda* performs the analysis in the application. *Quanteda* is a free package available through the Comprehensive R Archive Network (CRAN) that enables the creation of corpus tokens, dictionary analysis, and application of topic models (Benoit et al., 2018). In this study, we used the *dfm* function to refer to the dictionary and to count the number of words corresponding to each level. The dictionary used for this analysis was “ALL_sep” dictionary (comprised of all the parts of speech and all levels) included in the CEFR-J wordlist.

At this stage, this tool fails to distinguish between the same word with different parts of speech, and words with multiple parts of speech are counted multiple times. For example, the word “next!” can have three parts of speech: adverb (A2), determiner (A2), and pronoun (B2); this application is unable to differentiate between each part of speech. Therefore, “next” is counted for both the A2 and B2 levels, requiring us to manually categorize these words with multiple parts of speech.

Figure VI: Annotated Classroom Corpus Sample

```xml
<?xml version="1.0" encoding="Shift_JIS"?>
<body>
  <greetings>
    <teacher>
      <eng>
        <s>Good, morning, miss.○○</s>
      </eng>
    </teacher>
    <student>
      <eng>
        <s>Good, morning, Mr. Taguchi.</s>
      </eng>
    </student>
    <teacher>
      <eng>
        <s>How are you today?</s>
      </eng>
    </teacher>
    <student>
      <eng>
        <s>I’m fine, thank you, and you?</s>
      </eng>
    </student>
  </greetings>
</body>
```
Pilot Study

This brief section demonstrates how to utilize the Classroom Corpus Vocabulary Analyzer (CCVA) with the CEFR-J wordlist, followed by displaying the results and its implications. The following sub-sections describe the materials, results, and the provisional evaluation of the CCVA.

Materials

Five student teachers (juniors) contributed their English class transcripts for their teaching practicums at public junior high schools in Japan. Some taught 7th graders, and the others taught 8th graders. The transcripts were annotated with the speaker and language tags in XML format. The CCVA analyzed the transcripts with the path in order to extract the instructor's English language speech. Table 2 shows the participants’ profiles. L2 proficiency was measured based on the scores of the Cambridge Assessment General English (URL: https://www.cambridgeenglish.org/test-your-english/general-english/) using 25 multiple-choice questions. Grades 7 through 9 are included in lower secondary schools, and the classes generally consist of 40 students maximum.

<table>
<thead>
<tr>
<th>Participant ID</th>
<th>Grade</th>
<th>Target material items to teach</th>
<th>Number of instructor English tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (19/25)</td>
<td>7th</td>
<td>Third-person singular present tense</td>
<td>2,364</td>
</tr>
<tr>
<td>2 (15/25)</td>
<td>7th</td>
<td>Interrogative, <em>when, where, what time, how many, how long</em></td>
<td>1,288</td>
</tr>
<tr>
<td>3 (13/25)</td>
<td>8th</td>
<td>To-infinitive (adverbial)</td>
<td>1,168</td>
</tr>
<tr>
<td>4 (21/25)</td>
<td>9th</td>
<td>Present participle post modification</td>
<td>2,293</td>
</tr>
<tr>
<td>5 (25/25)</td>
<td>8th</td>
<td>To infinitive (adjectival)</td>
<td>1,078</td>
</tr>
</tbody>
</table>

RESULTS AND DISCUSSION

Figure 7 shows a sample CCVA analysis result. The CCVA analyzed the transcript via the linefeed and yields from the columns left to the right: the line number, the number of word types and tokens, the number of sentences in each line, the number of tokens for CEFR-J levels A1 to B2, followed by the text in the next column, and the tokens’ respective CEFR-J proportions in each line. We must note that the contractions, such as “I’m,” are delimited by the comma and counted as two separate tokens and types as well (doc_ids 5 and 6). However, the text in the doc_id 6, “m sleepy,” indicating that the CCVA analyzed “(I) am sleepy” as including only “sleepy” as in CEFR-J level A2, but not for “m” as “am” that should be analyzed, as in CEFR-J level A1.
The CCVA results can also be used to show the CEFR-J vocabulary distribution for each level (Figure 8). Generally, the student teachers’ use of the English language turned out to be the CEFR-J level A1, which is the fundamental level, attracted the most significant ratios (58.00–77.37%; x²(4) = 3.362, ns). A sharp drop occurred in Level A2 (8.67–15.44%; x²(4) = 3.412, ns), and the drop changed to a gradual decrease in B1 (x²(4) = 3.412, ns) and B2 (x²(4) = 2.545, ns), the final, relatively more advanced, and least frequently used word band. Although student teachers’ vocabulary usage showed no statistical differences in the ratios of each vocabulary level (A1, A2, B1, and B2), according to student teachers’ English proficiency levels, a variation in the usage of vocabulary occurs. The results shown by the CCVA help student teachers notice trends in their vocabulary usage, revealing the vocabulary items that they are not likely to use and the ones they overuse in the classroom, which provides the teacher with materials that help them improve their classes in terms of vocabulary usage.
The previous two sub-sections describe how to utilize the CCVA and the results of the pilot study. The English classroom transcript analysis results demonstrated that the student teachers’ use of English in their teaching practice lessons could be represented by the CEFR-J vocabulary levels. The majority of the English tokens observed at the most basic level was level A1. Although the words in the transcripts may have been counted multiple times for different levels, the word token ratio distributions on the CEFR-J four levels appeared to indicate the general tendency of the spoken language that McCarthy (1999) argued. In this respect, the CCVA has the potential to analyze English utterances spoken in the classroom, revealing insights on vocabulary usage.

This section focused on the pilot study by utilizing the CCVA in order to analyze the classroom transcripts. Displaying the results suggested a potential application for classroom corpus analysis, since the CCVA can classify tokens from the classroom transcripts based on the CEFR-J’s four-word levels (A1, A2, B1, and B2). The following section discusses the CCVA’s pedagogical implications, lists limitations, and suggests areas for future research.

**CONCLUSION**

We discussed the CEFR-J development that the CEFR inspired, and the CEFR-J-based Vocabulary Level Analyzer (Negishi & Uchida, 2019) that analyzes textual and grammatical
features. Based on these two main sources, we developed the CCVA (Classroom Corpus Vocabulary Analyzer with the CEFR-J wordlist). This section summarizes how we can utilize the CCVA, despite its limitations, and proposes future research.

The pilot study revealed that the CCVA interpreted the classroom transcript annotated with the speaker tags and language tags in XML format, and this resulted in the English tokens’ distribution across the CEFR-J’s four levels (A1, A2, B1, and B2). With these tags, we can designate the target speakers (i.e., teachers, students, or both). We can also pinpoint the language used in the pilot study, the English language. By specifying the speaker and what language is being analyzed, researchers can extract the words in the transcripts and analyze the word level distribution based on the CEFR-J’s four levels (A1, A2, B1, and B2).

Researchers, teacher trainers, and the teachers in the transcripts can utilize the word level distribution to develop the teachers’ classroom speech by controlling the difficulty of the classroom English utterances. The CCVA allows researchers to conduct research specific to the CEFR-J, in terms of vocabulary. Observing a variation of vocabulary levels in a classroom possibly facilitates teachers’ using a more comprehensive range of vocabulary items that pertain to different CEFR-J vocabulary levels.

We are aware of at least one limitation to the CCVA analysis. Since this tool fails to distinguish between words with multiple parts of speech, the findings require manual review. One way to address this issue is to annotate the transcripts with part-of-speech (POS) tags before conducting the CCVA analysis. We may also equip the CCVR with the POS-annotated CEFR-J wordlist. We will be able to avoid duplicate counting for tokens and types with multiple POS.

The CEFR-J has contributed to addressing the lack of a consistent system for measuring language learning proficiency. Applying the contextualized common framework of the CEFR-J to the original studies, the findings possibly reveal the current problems to tackle. We hope the usage of the CCVA encourages the publication of further research on the CEFR-J in the future with the expectation of the development of other descriptor-based tools to be created in Asia.

REFERENCES


