The Bulletin of International Education Center Tokai University
Vol. 1, No.1, 2019, pp. 1-21
DOI: 10.18995/24347337.1-1.1

Examining Japanese English Learners' Vocabulary Characteristics in a Paired Conversation: A Case Study of the CEFR B1 Speaking Test

Hiroko USAMI
Tokai University

Abstract

Vocabulary knowledge has been widely investigated from different perspectives (e.g., Read, 2000; Nation, 2001). In line with the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001), the English Vocabulary Profile (EVP) provides six CEFR levels with the individual meanings of each word and phrase based on the learners' writing. However, learners' vocabulary in their speech has not yet been sufficiently examined in terms of CEFR levels, on the basis of EVP. Therefore, the aim of this research was to examine Japanese English learners' overall spoken and CEFR vocabulary levels used in the CEFR B1 Speaking Test from the Cambridge English Qualifications B1 level, which makes use of corpus linguistic techniques such as wordlists, n-grams, and keywords. The results showed that nearly two-thirds of the participants obtained A1 in their overall spoken production, and the participants' CEFR vocabulary levels were slightly lower than their overall spoken levels. Approximately 70% or more of A1 level vocabulary was used by all CEFR level participants. However, when examining the participants' vocabulary in more detail, such as use of conjunctions, articles, modal verbs, interrogatives, and adverbs (which are relatively difficult to use in speech), participants with A2+, B1, B1+, and B2 CEFR levels were more likely to use this vocabulary.

Keywords: CEFR, vocabulary, speech, wordlist, keyword list

1. Introduction

Vocabulary knowledge has been widely examined and assessed from different perspectives (e.g., Read, 2000; Nation, 2001). In line with the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001), the linguistic features of learners' English competence have been examined as part of the English Profile Program (EPP) (Hawkins and Filipović, 2012). In the English Vocabulary Profile (EVP), one of the six CEFR levels is assigned to the individual meanings of each word and phrase. Rather than using speech, however, the assignment of CEFR level

information is based on samples of learners' writing contained in the learner corpora.

Using learner corpora, learners' vocabulary has been mainly researched as Contrastive Interlanguage Analysis (CIA), compared to native speakers' or other L2 learners' vocabulary (e.g., Granger, 1998; 2002). However, learners' vocabulary has not yet been examined adequately in terms of CEFR levels, on the basis of EVP. Therefore, the aim of this study was to examine the CEFR levels and characteristics of Japanese English learners' vocabulary in their speech, utilising corpus linguistic techniques such as wordlist, n-grams, and keywords.

In order to examine learners' CEFR vocabulary level, the CEFR B1 Speaking Test was administered, which is one of the speaking tasks from the Cambridge English Qualifications B1 level and involves a paired-conversation task while looking at an image. The vocabulary from this conversation was examined, with the rate of words reported for each CEFR vocabulary level, wordlist, n-gram, and keywords, as well as their overall spoken and vocabulary CEFR levels.

2. Literature Review

2.1 Research using Spoken Learner Corpus

Learners' speaking skills have been researched and assessed from different perspectives, such as pronunciation, vocabulary, grammar, and discourse (e.g., Luoma, 2004; Hugh, 2010). In corpus linguistic research, as part of the Contrastive Interlanguage Analysis (CIA), learners' speech has been analysed and examined in terms of their underuse or overuse of particular words or phrases, compared to that of different L2 learners or native speakers of English (e.g., Granger, 1998; 2002).

As a worldwide spoken (dialogue) learner corpus, the Louvain International Database of Spoken English Interlanguage (LINDSEI) (Gilquin, DeCock, and Granger, 2010) was composed of dialogues by 554 learners from 11 European countries and Japan. The International Corpus Network of Asian Learners of English (ICNALE) Spoken Dialogue is currently being constructed as a 3-year project from 2017 to 2019, followed by ICNALE Written Essays and ICNALE Spoken Monologue (Ishikawa, 2013; 2014; 2018). ICNALE Spoken Dialogue is developing a collection of Asian learners' oral interviews, including picture descriptions and role-plays.

The National Institute of Communication and Information Technology, Japanese Learners of English (NICT JLE) Corpus (Izumi, Uchimoto, and Isahara, 2004) is a relatively large spoken (dialogue) corpus of 1,281 Japanese learners' oral interviews based on the Standard Speaking Test (SST). Using this corpus, Tono (2004) examined the productive vocabulary of learners in terms of wordlist, structure, and collocation across

their SST proficiency levels. Other types of English proficiency levels such as TOEIC® and TOEFL®[1] are provided. However, Japanese learners' CEFR levels based on tasks from Cambridge English Qualifications tests are not tagged in the above spoken corpora.

2.2 CEFR Vocabulary in Learners' Speech and Writing

Vocabulary in both learners' spoken and written formats has been researched in the context of CEFR. As for research on learners' vocabulary in their writing, Chen and Baker (2016) examined lexical bundles of learners' essays across their CEFR B1, B2, and C1 levels. Leńko-Szymańska (2015) examined text length and lexical characteristics based on the EVP in learners' essays of IELTS.[2] Barker (2015) presented data for each CEFR level and investigated different samples of L2 learners' writing on each of the six CEFR levels in terms of EVP, the English Grammar Profile (EGP; Hawkins and Filipović, 2012), and functions (Green, 2012). The learners' CEFR writing levels almost matched their CEFR vocabulary levels; A1 and A2 learners used A1- and A2-level vocabulary, B1 and B2 learners used vocabulary ranging from A1 to B2, and C1 and C2 learners used all CEFR level vocabulary. Focusing on Japanese learners, Usami (2018) examined the Japanese learners' CEFR productive vocabulary level in their essays, compared to their CEFR receptive vocabulary knowledge. Contrary to the results described in Barker's (2015) study, more than 70% of the vocabulary used by learners at all CEFR levels was A1 level, although their overall writing CEFR levels were A2 and B1.

Khalifa and Salamoura (2011) pointed out that spoken corpora have not been used in validating speaking tests. As for research on learners' vocabulary in their speech, Galaczi (2003) examined the interaction patterns in First Certificate Exam speaking tests. Hulstijn, Schoonen, De Jong, Steinel, and Florijn (2011) examined the speech of Dutch adult English learners in terms of their productive vocabulary and grammar knowledge, speed, and pronunciation. Learners' speech at B1 and B2 levels were discriminated well by their productive vocabulary and grammar on paper-based tests. Usami (2016) investigated Japanese learners' CEFR vocabulary in paired conversations across two different topics. Their speaking CEFR level ranged from A1+ to A2+, and they used almost 50% of A1 and almost 10% of A2 level vocabulary, along with many fillers and Japanese words. In Usami (2019), Japanese English learners' receptive and productive vocabulary was compared using the CEFR B1 Speaking Test, which was also used in this

[1] The Test of English for International Communication (TOEIC) is an English proficiency test used in the global work environment, and the Test of English as a Foreign Language (TOEFL) is an English proficiency test for study, immigration, and work. See https://www.ets.org/toeic/ and https://www.ets.org/toefl/ for TOEIC® and TOEFL®, respectively.

[2] International English Language Testing System (IELTS) is an English proficiency test for higher education and global migration. See IELTS: https://takeielts.britishcouncil.org.

research. Their results showed that their receptive vocabulary level demonstrated on the multiple-choice vocabulary test was high, whereas their productive CEFR vocabulary level was low, with approximately 60% of the participants at the A1 level.

While there has been prior research on learners' speech vocabulary in the context of CEFR (e.g., Galaczi, 2003; Hulstijn et al, 2011), the characteristics and frequencies of the vocabulary used in their speech have not been deeply analysed in the CEFR context. This study, therefore, aimed to examine the characteristics of Japanese English learners' vocabulary in the CEFR context, using a paired-conversation task and presenting their wordlists, n-grams, and keywords according to their CEFR levels. The following research question was examined: What are the characteristics of Japanese English learners' vocabulary in a paired conversation based on their CEFR levels?

3. Method

3.1 Task

In this study, a paired, learner-learner conversation task was administered to 165[3] university students. Focusing on their speaking CEFR levels, a task called 'Summer Job' was selected from one of the past Cambridge English Qualifications B1 Preliminary speaking tests (CEFR B1 Speaking Test). During the task, an image (see Figure 1) was given to each pair of randomly selected participants, followed by instructions in both English and Japanese. An examiner said to the participants, 'I'm going to describe a situation to you. A young man is going to travel to England to do a summer job. Talk together about the different things he should take with him and say which would be the most useful', first in English and then in Japanese. Participants were then allowed 30 seconds for preparing their conversation individually and were given English and Japanese instructions by an examiner again. Finally, they were required to start by stating their name and a short greeting and then continue their conversations for two minutes, looking at the image.

[3] In this study, there were 166 students participating in 83 pairs. However, the transcription of one student was not used because of incomplete data.

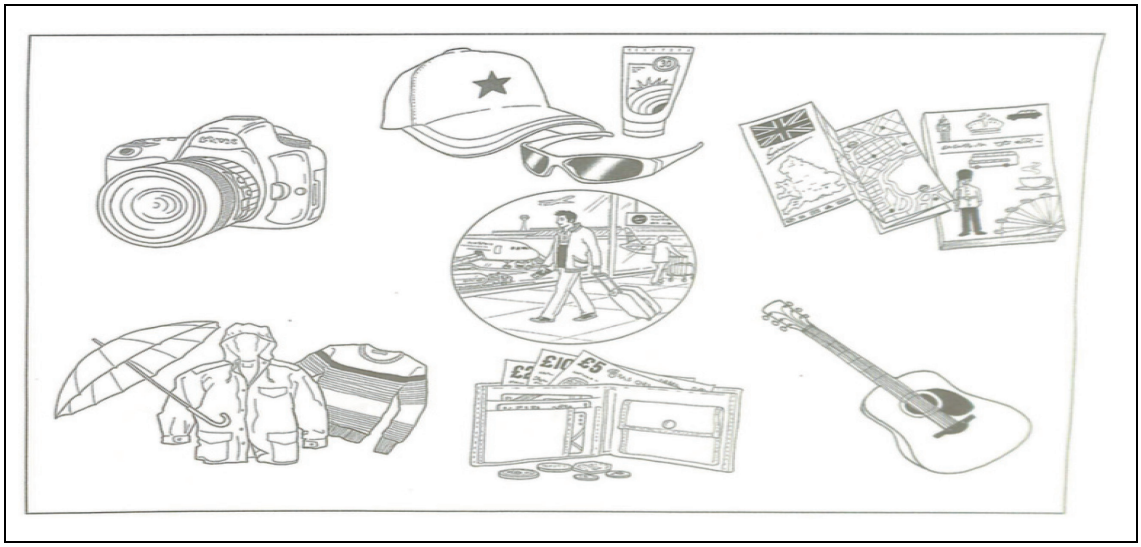

Figure 1. The image for the task called 'Summer Job' obtained from Cambridge English (2014: v).

3.2 Participants

In this study, the task was administered to 165 university students (154 first-year, 5 second-year, 3 third-year, and 3 fourth-year students). As shown in Table 1, most participants (93.3%) were first-year students.

Table 1.

Year level of participants

Year level	1st	2nd	3rd	4th	Total
# (%)	154 (93.3%)	5 (3.0%)	3 (1.8%)	3 (1.8%)	165 (100%)

3.3 CEFR Learner Corpus

The conversations of all pairs were audio-recorded, transcribed, and stored as one file, along with separate files for each speaker, into a CEFR Learner Corpus created by the author. The CEFR Learner Corpus is considered to be a monitor corpus, storing Japanese learners' essays and speeches on tasks selected from Cambridge English Qualifications in the future. The CEFR Learner Corpus will be utilised, focusing on examining the learners' writing and speaking CEFR levels based on tasks from past Cambridge English Qualifications tests.

Their transcribed speech in the CEFR Learner Corpus was annotated with various information, including metadata such as target skills, target CEFR levels of the task, tasks, time, topics, and test conditions. Participants' data such as year level, major, class, and class level were also captured.

In addition, a professional CEFR rater rated each participant's conversation according to 13 CEFR levels (Pre-A1, A1, A1+, A2, A2+, B1, B1+, B2, B2+, C1, C1+, C2, and C2+) in terms of three categories: 1) overall spoken production, 2) vocabulary range and control, and 3) grammatical accuracy. Furthermore, each participant's transcription was analysed in terms of statistics such as type, token, and type–token ratio. Also, the percentage of each CEFR vocabulary level used in their conversations, obtained on the website Text Inspector (see Figure 2),[4] was added.

Word List	Types	Tokens
A1	32 (52.46%)	53 (59.55%)
A2	12 (19.67%)	19 (21.35%)
B1	6 (9.84%)	6 (6.74%)
B2	5 (8.20%)	5 (5.62%)
C1	1 (1.64%)	1 (1.12%)
Unlisted	5 (8.20%)	5 (5.62%)

Figure 2. An example of each CEFR vocabulary level on the Text Inspector

3.4 Corpus Analytical Tool

In this research, in order to examine the characteristics of the vocabulary used by the participants in more detail, a software called My Dictionaries in Wmatrix[5] was used. My Dictionaries was developed and will be further improved in collaboration with Paul Rayson at Lancaster University, UK. My Dictionaries was developed to compensate for the limitations of Text Inspector and provides more detailed corpus analytical techniques such as concordance, cluster, wordlist, n-grams, and keyword list, as well as statistics such as type, token, and type-token ratio. My Dictionaries was utilised in this study to examine the vocabulary used in learners' speech, based on the details of each CEFR level of the learner group.

[4] See Text Inspector website: https://textinspector.com.
[5] See Wmatrix website: http://ucrel.lancs.ac.uk/wmatrix/.

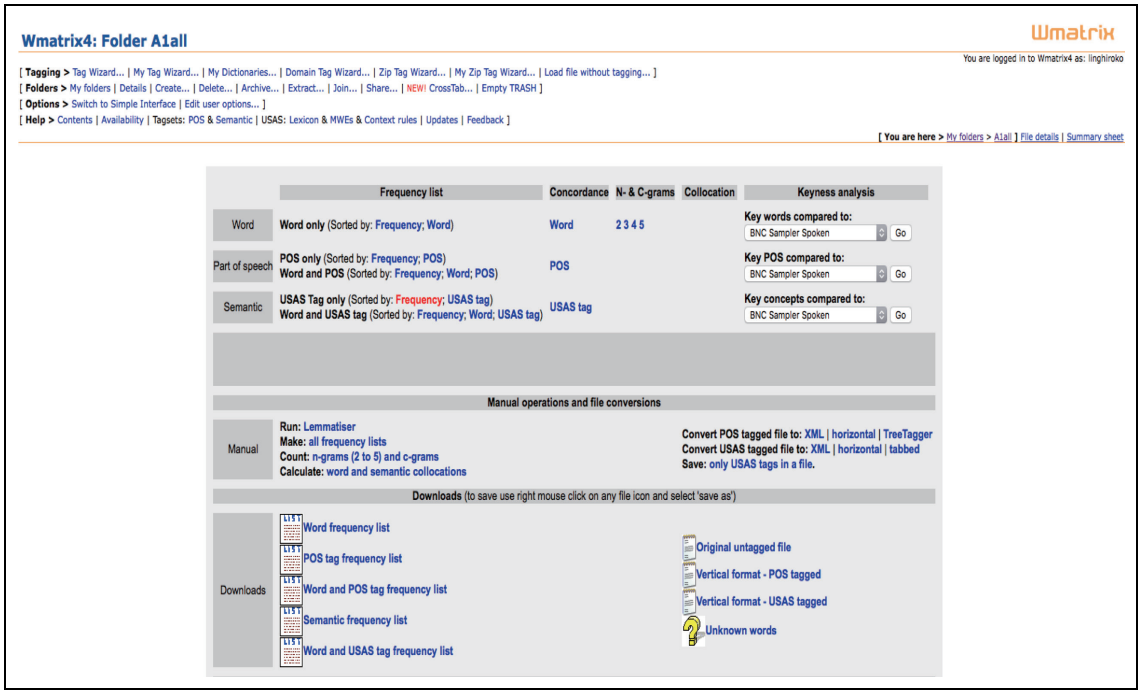

Figure 3. An example of the analysis of the A1 learner group in My Dictionaries

4. Results and Discussion

4.1 CEFR Levels of Speech and Vocabulary

First, each participant's CEFR level on the task 'Summer Job' for 'overall spoken production', one of the CEFR level categories that is rated by the professional CEFR rater mentioned above, was examined.

Table 2.

Participants' CEFR levels on 'Summer Job' for overall spoken production

Pre-A1	A1		A2		B1		B2	
13 7.9%	105 63.6%		35 21.2%		10 6.1%		2 1.2%	
	A1	A1+	A2	A2+	B1	B1+	B2	B2+
	80 48.5%	25 15.2%	27 16.4%	8 4.8%	9 5.5%	1 0.6%	2 1.2%	0 0.0%

Table 2 indicates the number and percentage of participants across each CEFR level. Each participant's speech was rated on the basis of 13 CEFR levels. Besides these, each participant's CEFR level was categorised according to broader, combined categories such as A1 (A1 and A1+), A2 (A2 and A2+), B1 (B1 and B1+), and B2 (B2 and B2+).

According to Table 2, nearly two-thirds of the participants (63.6%) were categorised as A1 (A1 and A1+), followed by A2 (21.2%) (A2 and A2+). Very few participants were categorised as B1 (6.1%) (B1 and B1+) and B2 (1.2%) (B2 and B2+). When analysed in more detail, almost half of the participants were categorised as A1 (48.5%), followed by A2 (16.4%) and A1+ (15.2%). Few participants were categorised as Pre-A1 (7.9%), B1 (5.5%), and A2+ (4.8%), and very few were categorised as B2 (1.2%) and B1+ (0.6%). The paired-conversation task administered in this research targeted participants with B1 CEFR level. However, the actual CEFR levels of most participants were below B1.

Once each participant's CEFR level for overall spoken production was obtained, their CEFR levels for vocabulary range and control were analysed.

Table 3.

Participants' CEFR levels on 'Summer Job' for vocabulary range and control

Pre-A1	A1		A2		B1		B2	
24 14.5%	94 56.9%		38 23.0%		8 4.8%		1 0.6%	
	A1	A1+	A2	A2+	B1	B1+	B2	B2+
	81 49.1%	13 7.9%	23 13.9%	15 9.1%	6 3.6%	2 1.2%	1 0.6%	0 0.0%

Table 3 indicates the number and percentage of participants across each CEFR level. According to Table 3, more than half of the participants (56.9%) were categorised as A1 (A1 and A1+), followed by A2 (23.0%) (A2 and A2+). Compared to their CEFR levels of overall spoken production indicated in Table 2, fewer participants were categorised as A1, whereas a few more participants were categorised as A2. Very few participants were categorised as B1 (4.8%) (B1 and B1+) and B2 (0.6%) (B2 and B2+). Compared to the result in Table 2, fewer participants were categorised as both B1 and B2. When analysed in more detail, almost half of the participants were categorised as A1 (49.1%), followed by Pre-A1 (14.5%) and A2 (13.9%). Few participants were categorised as A2+ (9.1%), A1+ (7.9%), and B1 (3.6%), and very few were categorised as B1+ (1.2%) and B2 (0.6%). Therefore, compared to their CEFR levels of overall spoken production, as indicated in Table 2, their CEFR vocabulary levels seem to be lower; 6.6% more of the participants were categorised as Pre-A1 for their vocabulary, though 4.3% more of the participants were categorised as A2+ for their vocabulary.

As illustrated above in Tables 2 and 3, participants' CEFR levels for overall spoken

production do not seem to match those of vocabulary range and control. Therefore, participants' CEFR overall spoken production levels that were categorised as per the CEFR level for their vocabulary range and control was examined, using SPSS.

Table 4 indicates participants' CEFR levels across their overall spoken production as well as vocabulary range and control, while Figure 4 indicates whether their overall spoken production CEFR levels are higher or lower than their vocabulary range control CEFR levels.

Table 4.

Participants' CEFR levels across their overall spoken production and vocabulary range and control

V / O	Pre-A1	A1	A1+	A2	A2+	B1	B1+	B2
Pre-A1	61.5%	20.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
A1	38.5%	75.0%	48.0%	14.8%	0.0%	0.0%	0.0%	0.0%
A1+	0.0%	2.5%	32.0%	11.1%	0.0%	0.0%	0.0%	0.0%
A2	0.0%	2.5%	16.0%	44.4%	37.5%	22.2%	0.0%	0.0%
A2+	0.0%	0.0%	4.0%	25.9%	62.5%	22.2%	0.0%	0.0%
B1	0.0%	0.0%	0.0%	3.7%	0.0%	44.4%	100.0%	0.0%
B1+	0.0%	0.0%	0.0%	0.0%	0.0%	11.1%	0.0%	50.0%
B2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	50.0%

Note. V = vocabulary range and control. O = overall spoken production.

As shown in Figure 4, Pre-A1 participants stayed at the same CEFR level or higher. As their CEFR levels increase, A1 level participants were classified as lower CEFR levels, although A1+ and A2 level participants remained relatively evenly distributed among those who stayed at the same CEFR level and those who were classified as higher or lower CEFR levels. A2+, B1, and B2 level participants tended to stay at the same CEFR level or lower, although approximately 10% of B1 level participants were classified as higher CEFR levels. B1+ level participants stayed at the lower CEFR level.

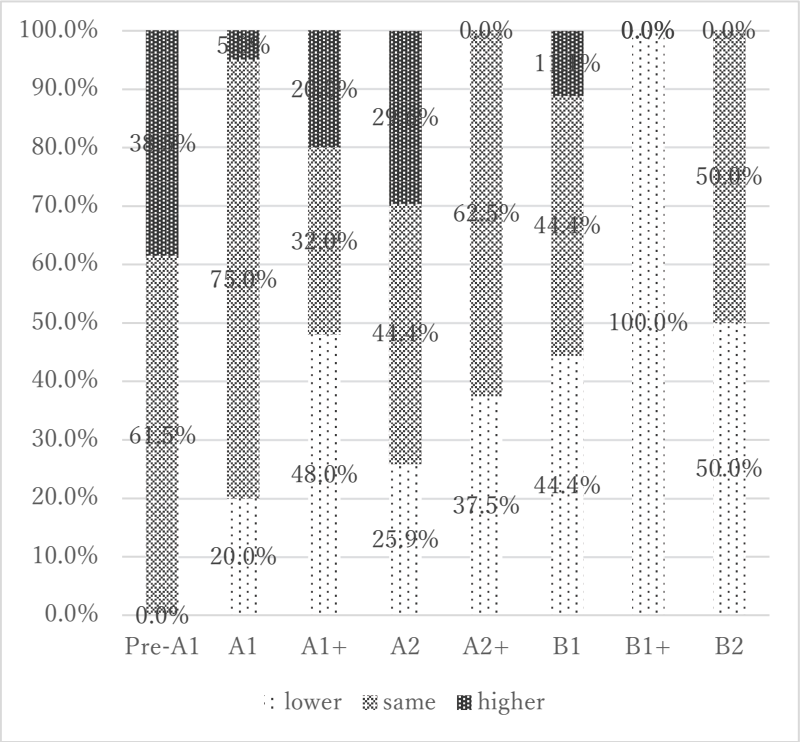

Figure 4. Comparison of participants' CEFR levels across spoken and vocabulary

4.2 Type, Token, and Type–Token Ratio

Figure 5 indicates the average achievement by type, token, and type–token ratio (TTR) of each CEFR participant's conversation. Each participant was required to discuss the topic for two minutes; however, each type, token, and TTR varied across their CEFR levels. Generally, participants with higher levels in speech tend to speak more (e.g., Tono, 2004; Ishikawa, 2008), as shown by the higher number of produced tokens. In this study, there were some exceptions, especially among B1 and B1+ participants, who tended to speak less than the lower A2+ participants and speak much less than the higher B2 participants. This is in contrast with the results of the analysis of SST data by Tono (2004) and Ishikawa (2008). Furthermore, participants with A2+ and B2 CEFR levels tended to use a more varied vocabulary, which was in line with the findings of Tono (2004) and Ishikawa (2008), despite B1 and B1+ participants using a slightly less varied vocabulary. B2 participants spoke the most of all the groups and used the most varied vocabulary.

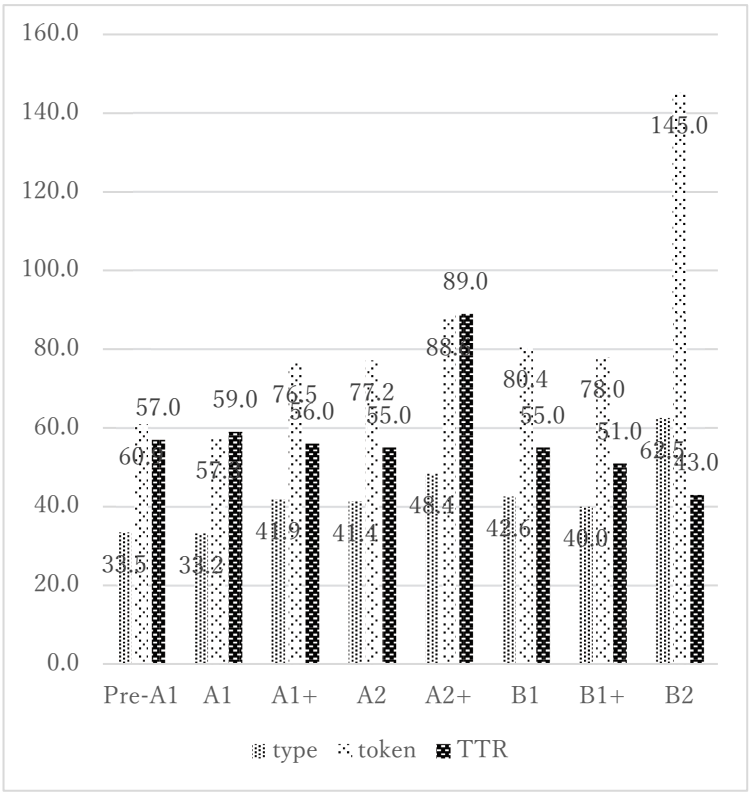

Figure 5. Type, token and TTR across CEFR levels of their overall spoken production

4.3 CEFR Levels of Vocabulary Used in Learners' Conversation

In this section, the extent to which participants of each CEFR level in overall spoken production actually used CEFR level vocabulary in their paired conversation as examined using Text Inspector is discussed.

Table 5 indicates the percentage of each CEFR level used in their conversation across CEFR levels for overall spoken production. It was apparent that participants of all CEFR levels used approximately 70% or more of A1 level vocabulary, whereas very few participants used vocabulary of the B2 level or higher, which were results also found by Usami (2019). Furthermore, for the participants in the Pre-A1 to A2 level range, the higher their CEFR levels, the more A1 level vocabulary they tended to use (from 69.08% to 75.37%), whereas participants above the A2+ level (i.e., at higher CEFR levels) used less A1 level vocabulary (from 72.95% to 72.46%). Conversely, participants above the A2+ level tended to use higher-level vocabulary (from 16.23% to 22.46%). There was an exception, however, as A1+, A2, and A2+ participants used C1 level vocabulary frequently. Interestingly, UL, which is unlisted on the EVP, was used most frequently by

Pre-A1 participants and the least by B2 participants. In the following section, the type of words used across each CEFR level participant is examined.

Table 5.

CEFR levels of vocabulary used in their paired conversation across their overall spoken production CEFR levels

	Pre-A1	A1	A1+	A2	A2+	B1	B1+	B2
A1	69.08%	74.27%	74.02%	75.37%	72.95%	70.35%	74.01%	72.46%
A2	17.29%	15.80%	16.21%	15.92%	16.23%	18.44%	17.68%	22.46%
B1	1.82%	1.61%	1.18%	1.36%	1.30%	1.20%	2.02%	0.72%
B2	0.08%	0.47%	0.22%	0.20%	0.39%	1.81%	0.00%	0.00%
C1	0.08%	0.09%	0.20%	0.21%	0.23%	0.00%	0.00%	0.00%
C2	0.00%	0.01%	0.06%	0.04%	0.00%	0.00%	0.00%	0.00%
UL	11.63%	7.72%	8.09%	6.88%	8.86%	8.18%	6.27%	4.35%

4.4 Wordlist and Keyword List

The kind of word that is actually used by each CEFR level participant was examined by analysing their wordlist and keyword list. Table 6 illustrates the top 20 most frequently used words by each CEFR level participant.

First, as for personal pronouns, participants with Pre-A1 to A2 CEFR levels used the first person pronoun, *I*, most frequently, whereas participants with A2+ to B2 CEFR levels used the third person pronoun, *he*, most frequently to refer to the man in the image shown in Figure 1. As for verbs, *think* was frequently used by all CEFR level participants, probably because the phrase "*I think"* was frequently used. Also, the Be-verb *is* appeared frequently across all CEFR level participants, and participants with A2+, B1, B1+, and B2 CEFR levels could also use various verbs such as *bring*, *take*, *travel*, *have*, *need*, and *know*.

Table 6.

Wordlist across CEFR levels of their overall spoken production

R	Pre-A1	A1	A1+	A2	A2+	B1	B1+	B2
1	I	I	I	I	he	he	he	he
2	yeah	think	think	he	I	I	okay	you
3	is	is	is	think	think	is	I	think
4	think	he	he	is	is	think	think	I
5	oh	you	the	so	and	you	should	is
6	he	money	so	you	do	if	is	bring
7	yes	so	to	do	yes	the	may	and
8	you	do	and	and	take	do	name	so
9	money	the	yes	yeah	a	and	bring	do
10	and	yes	money	the	n't	to	guide	a
11	important	to	you	money	so	because	because	yeah
12	do	yeah	useful	to	yeah	so	where	know
13	hmm	and	oh	England	we	useful	so	wallet
14	map	useful	do	important	the	n't	hello	have
15	if	because	yeah	because	important	can	my	to
16	hi	camera	because	n't	this	bring	what	need
17	okay	most	guitar	if	should	thing	your	why
18	useful	map	a	can	bring	travel	map	probably
19	we	England	map	what	if	have		it
20	most	if	this	map	but	this		UK

Note. R = Rank.

Second, different conjunctions could be used by all CEFR level participants; however, participants with B1 and B1+ CEFR levels tended to use *because* and *if* frequently, both of which were being used following a sentence. The use of articles is difficult for Japanese English learners, as discussed in Izumi and Isahara (2004), but even participants with A1 to B1 CEFR levels seemed able to use the definite article *the*, and the indefinite article *a* seemed to be used frequently by A2+ and B2 participants, though A1+ participants did use it sometimes as well.

Third, modal verbs such as *should*, *can*, and *may*, and the interrogatives *what*, *where*, and *why* were used mostly by participants with A2 to B2 CEFR levels. In addition, the adverb *probably* was used only by B2 participants, while fillers such as *yeah*, *oh*, *hmm*, and *okay* were frequently used by Pre-A1 participants.

Finally, as for the use of nouns, which are required to answer the question of this task properly (illustrated in Figure 1), *money* was used frequently by participants with Pre-A1 to A2 CEFR levels, which would indicate that *money* was apparently the answer to the question of what would be most useful for the man to take with him to England for his summer job. Participants with B1+ and B2 CEFR levels used nouns such as *guide*, *wallet,* and *map* in their responses.

Table 7.

Keyword list across CEFR levels of their overall spoken production

R	Pre-A1	A1	A1+	A2	A2+	B1	B1+	B2
1	yeah	money	London	he	take	if	okay	probably
2	carry	useful	or	will	pictures	during	may	UK
3	hmm	yes	buying	we	sunscreen	sun	should	guide
4	oh	ah	its	pictures	we	make	needed	wallet
5	bag	most	too	be	cameras	large	guide	record
6	where	too	play	England	heavy	kingdom	map	bad
7	item	necessary	oh	know	a	he	where	bring
8	hi	camera	helps	n't	lets	travel	rains	might
9	travel	that	forgot	traveling	but	needed	he	everywhere
10	enjoy	oil	view	foreign	next	than	else	does

Note. R = Rank.

Table 7 indicates the keyword list across CEFR levels of participants' overall spoken production. First, the personal pronouns *I* and *he*, and verbs *think* and *is* are not keywords, probably because all of these are used frequently across all CEFR levels of participants. Instead, verbs other than *think* and *is* are actually keywords. Relatively easy and simple verbs such as *carry*, *travel*, and *enjoy* were used by Pre-A1, and different verb forms such as *–ing* (*buying* and *traveling*), *-s* (*helps*), and the past tense form (*forgot*) were used by

A1+ and A2 participants (e.g., *money helps everything* and *I forgot the map*). B1 and B1+ participants used *needed*, which is a past participle form (e.g., *guitar is not needed*).

Furthermore, conjunctions and articles did not seem to be keywords either. However, B1 participants used *than*, which is used in comparisons, and a preposition *during* (e.g., *during the travel*). Modal verbs such as *may*, *should*, and *might* were keywords for B1+ and B2 participants (e.g., *you might think he needs a camera*) and the adverb *probably* (e.g., *probably he should bring a wallet*) was the rank 1 keyword among B2 participants. Fillers were keywords for Pre-A1 participants and terms such as *yeah*, *hmm*, and *oh* were often used.

Finally, regarding the nouns used in the answer to the question, A1 participants tended to answer *money* and sometimes *camera*, both of which were actually shown in the image (Figure 1). A2+ participants mentioned *sunscreen* and *cameras*, both of which were shown in the image as well. However, *pictures*, which was related to *camera*, was also frequently mentioned by A2 and A2+ participants. B1+ participants mentioned *map*, which was actually present in the image. However, B1+ and B2 participants also used *guide*, instead of *map*, and *wallet*, instead of *money*.

4.5 N-gram List

In the previous section, one wordlist and keyword list are analysed and discussed across the participants' CEFR levels. This section considers slightly longer phrases to examine how each word and keyword were actually used in their conversation. Tables 8 and 9 indicate 2-gram and 3-gram, respectively, across the participants' CEFR levels of their overall spoken production.

First, regarding the combination of personal pronouns and verbs, the phrase *I think* was quite frequent across all CEFR level participants, as expected. Related to the use of *think*, participants with A1 to A2 CEFR levels used *think so*, whereas participants with A2+, B1+, and B2 CEFR levels used *think he*. According to Table 9, participants with A1, A1+, and A2 CEFR levels used *I think so* and *think so too*, whereas *I think he* was only used by participants with A2+, B1, and B1+ CEFR levels, and they did not use *I think so* or *think so too*.

Table 8.

2-gram across CEFR levels of their overall spoken production

R	Pre-A1	A1	A1+	A2	A2+	B1	B1+	B2
1	yeah yeah	I think	I think	I think	I think	I think	I think	I think
2	I think	do you	I' m	don' t	don' t	don' t	he should	you know
3	hmm hmm	I' m	do you	do you	he he	do you	think he	have to
4	I I	money is	is very	think he	take a	he can		to bring
5	don' t	think so	yes yes	yeah yeah	think he	you don'		he should
6	I don'	yes yes	money is	think so	have to	he he		he needs
7	oh oh	name is	so I	he he	he should	you think		does he
8	I' m	my name	it is	he can	I don'	he should		you have
9	is important	you think	you think	I' m	to bring	if if		think he
10	do you	he he	think so	so I	I' m	I' m		how many

Note. R = Rank.

The phrase *do you* seemed to be frequent across almost all CEFR levels and *you think* was also used though it was not so frequent. Table 9 indicates that *do you think* was relatively frequent across all CEFR levels. Interestingly, Table 8 shows that only B2 participants used *he needs* and *does he*, in which the *–s,* indicating third person pronouns, was used correctly. According to Table 9, *does he have* and *think he needs* were used by B2 participants, whereas A2 and A2+ participants used *he don't*, which is not correct in terms of the third person pronoun *–s*.

Participants with CEFR above A2 could use modals such as *he can*, *have to*, and *he should*. Regarding fillers, repeated fillers such as *yeah yeah*, *hmm hmm*, and *yes yes* were frequent across participants with Pre-A1 to A1+ CEFR levels. Moreover, Pre-A1 and A1 participants frequently used phrases related to their greeting (e.g., *my name is*, *hi I'm*, *nice to meet*) before they started to discuss.

Table 9.

3-gram across CEFR levels of their overall spoken production

R	Pre-A1	A1	A1+	A2
1	yeah yeah yeah	I think so	do you think	I think he
2	hmm hmm hmm	my name is	I think so	do you think
3	I don' t	do you think	so I think	what do you
4	my name is	I think he	my name is	I think so
5	hi I' m	what do you	I think the	my name is
6	hi my name	nice to meet	hi I' m	yeah yeah yeah
7	do you think	to meet you	what do you	don' t know
8	okay I think	how about you	I think money	think so too
9	is most important	think so too	think he should	he don' t
10	oh yeah yeah	I think the	is very important	I I think
R	A2+	B1	B1+	B2
1	I don' t	do you think	I think he	have to bring
2	I think he	you don' t		does he have
3	think he should	I think he		think he needs
4	don' t know	the most useful		you have to
5	we have to	useful thing is		
6	don' t have	most useful thing		
7	he don' t	a lot of		
8	is not important	don' t know		
9	do you think	I don' t		
10	yes I think	and he he		

Note. R = Rank.

Table 9 shows that A1+ and A2+ participants used adjectival phrases such as *is very important* and *is not important*. However, B1 participants used the superlative form such as *the most useful* and *most useful thing*.

5. Conclusion

In this study, the characteristics of Japanese English learners' vocabulary used in a

B1 level paired conversation task were examined using corpus linguistic techniques according to their CEFR levels, as well as their overall spoken and vocabulary CEFR levels.

Examining the participants' CEFR levels in overall spoken production, nearly two-thirds of the participants obtained A1 (A1 and A1+), and approximately 20% obtained A2 (A2 and A2+). Focusing on their CEFR levels in terms of their vocabulary range and control, their CEFR vocabulary levels seemed to be lower than their overall spoken CEFR levels. Therefore, vocabulary matching their spoken CEFR levels did not seem to be used enough in their speech. Even participants with A2+, B1, B1+, and B2 CEFR levels could use the same or lower CEFR levels of vocabulary, although A1+ and A2 participants seemed to try to use higher CEFR level vocabulary.

Examining which CEFR level of vocabulary was actually used across participants' CEFR overall spoken production levels, all participants used 70% or more of A1 vocabulary. However, participants with CEFR levels above A2+ used less A1 level vocabulary than other groups; for all participants with above A2+ level, the higher their CEFR levels, the more A2 level vocabulary they tended to use. In addition, UL (unlisted words on the EVP) were frequently used by Pre-A1.

Corpus linguistic techniques can reveal the characteristics of vocabulary and phrases used by the participants. Participants with above A2 CEFR levels could use conjunctions such as *because* and *if*, interrogatives, modals, adverb, a passive form, a third person singular *–s*, and superlatives, all of which seemed to be difficult to use in their speech. Conversely, participants with below A2 CEFR levels frequently used fixed expressions, such as *I think so too*, and greetings. All the participants tended to use *I think* quite frequently.

6. Limitations and Future Research

This research was conducted based on very limited data in terms of the number of participants and the task. This study reveals that the participants' CEFR levels of their vocabulary range and control were slightly lower than those of their overall spoken production, as also analysed in Usami (2019). Participants with above A2 CEFR levels could use relatively complicated grammatical structures though they did not use a higher level of vocabulary. Therefore, for future research, a task to analyse participants' grammar

in detail can be used to reveal participants' grammar proficiency level.

Acknowledgements

The present study was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number (K6K016). The present study was conducted under the approval of the Tokai University ethical committee (IRB No. 16032 approved in 2016 and 2017).

References

Barker, F. (2015). Portraits of each CEFL level. In J. Harrison & F. Barker (Eds.), *English Profile Studies* 5 (pp. 69-92). Cambridge: Cambridge University Press.

Cambridge English (May 26, 2014). *Cambridge English Preliminary 8 Preliminary English Test with answers*. Cambridge: Cambridge University Press.

Chen, Y.H., & Baker, P. (2016). Investigating critical discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics*, *37*, 6, 849-880.

Council of Europe. (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Galaczi, E. (2003). Interaction in a paired speaking test: the case of the First Certificate in English. *Research Notes 14*: 19-23.

Gilquin, G., DeCock, S., & Granger, S. (2010). Louvajn International database of spoken English Interlanguage. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Granger, S. (Ed.) (1998). *Learner English on Computer*. London: Longman.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3-33). Amsterdam and Philadelphia: John Benjamins.

Green, A. (2012). *Language functions revisited, English Profile Studies 2*. Cambridge: Cambridge University Press.

Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the common European framework, English Profile Studies 1*. Cambridge: Cambridge University Press.

Hugh, R. (2010). Teaching and Researching: Speaking. London: Routledge.

Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P., & Florijn, A. (2011). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, *29*(2), 203-221.

Ishikawa, S. (2008). *Eigo Corpus to Gengo Kyoiku* [*English Corpus and Language Education*]. Tokyo, Japan: Taishukan Shoten.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 1* (pp.91-118). Kobe, Japan: Kobe University.

Ishikawa, S. (2014). Design of the ICNALE Spoken: A new database for multi-modal contrastive interlanguage analysis. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 2* (pp. 63-76). Kobe, Japan: Kobe University.

Ishikawa, S. (2018). The ICNALE Spoken Dialogue no sekkei: Taiwa ni okeru L2 koto sanshutsu kenkyu notame ni [The design of the ICNALE Spoken Dialogue: For studies of L2 oral production in dialogues). In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 3* (pp.9-26). Kobe, Japan: Kobe University.

Izumi, E., & Isahara, H. (2004). Nihonjin eigo gakusyusya no eigo kanshi syutoku keikou no bunseki [Analysis of the acquisition of the English articles by Japanese English learners]. In E. Izumi., K. Uchimoto, & H. Isahara (Eds.), *Nihonjin 1200 nin no eigo speaking corpus* [English Speaking Corpus of 1200 Japanese people] (pp. 131-139). Tokyo: ALC.

Izumi, E., Uchimoto, K., & Isahara, H. (2004). *Nihonjin 1200 nin no eigo speaking corpus* [English Speaking Corpus of 1200 Japanese people]. Tokyo: ALC.

Khalifa, H., & Salamoura, A. (2011). Criterion-related validity. In L. Taylor (Ed.), *Examining Speaking: Research and Practice in Assessing Second Language Speaking, Studies in Language Testing volume 30* (pp. 259-292). Cambridge: UCLES/Cambridge University Press.

Leńko-Szymańska, A. (2015). The English vocabulary profile as a benchmark for assigning levels to learner corpus data. In M. Callies, & S. Götz (Eds.), *Learner corpora in language testing and assessment* (pp. 115-140). Amsterdam: John Benjamins Publishing Company.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

Nation, P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Tono, Y. (2004). The NICT JLE Corpus ni miru eigo gakusyusya no happyo goi no shiyo jyokyo [Trends in English learners' use in productive vocabulary seen in the NICT JLE Corpus]. In E. Izumi., K. Uchimoto, & H. Isahara (Eds.), *Nihonjin 1200 nin no eigo speaking corpus* [English Speaking Corpus of 1200 Japanese people] (pp. 96-112). Tokyo: ALC.

Usami, H. (2016). Vocabulary used in paired conversation across CEFR levels on the English Vocabulary Profile. *The Report of Foreign Language Center,* Tokai University, *36*, 39-48.

Usami, H. (2018). CEFR receptive and productive vocabulary knowledge of Japanese English learners. *Proceedings of APCLC 2018 Conference*, 480-484.

Usami, H. (2019). Japanese English Learners' Receptive and Productive Vocabulary Knowledge in Their CEFR B1 Speaking Test. *The Bulletin of International Education Center*, Tokai University, *39*, 23-35.