Does Vocalization Enhance L2 Formulaic Sequence Learning?

Hashizaki Ryotaro

Graduate School of Humanities, Nagoya University

Abstract

In foreign or second language (L2) classrooms, learners often vocalize when learning new words. Indeed, previous studies have demonstrated that vocalization is effective for memorization and this is called the production effect (PE). Furthermore, PE has been observed in the learning of L2 word meanings as well. However, whether this can be applied to larger language items is unknown. Thus, the present study tested the effect of vocalization on learning formulaic sequence (FS), a sequence of words such as collocations. To achieve this, the participants in the present study learned 90 verb-noun collocations and were tested on recognition of their forms and meanings both immediately and one week after the learning session. The result showed that vocalization enhances learning of forms and meanings of L2 FSs, and this effect is persistent over a one-week delay. This implies that PE is not affected by the nature of learned items and can be applied to larger units including FS such as collocations.

Does Vocalization Enhance L2 Formulaic Sequence Learning?

1. Introduction

When greeting others, we can use phrases such as "How are you?" and "How have you been doing?" so effortlessly that only a moment is needed to produce the phrases. This kind of ready-made language is termed as formulaic language, and when one refers to specific items, it is called formulaic sequence (Conklin & Schmitt, 2012) (FS). Wray and Perkins (2000) define formulaic sequence as:

> A sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar (p.1).

FS has been treated as an umbrella term for more than 40 kinds of items (Wray & Perkins, 2000), and with the growing frequency of research on the formulaic nature of language, the importance of learning and teaching FSs has also gained attention. The most dominant reason why FSs matter is due to their ubiquity in native speakers' output (Conklin & Schmitt, 2012). Conklin and Schmitt (2012) state that between one-third to even half of the entire output of native English speakers is formulaic. However, L2 speakers tend to focus on individual words rather than FSs (Arnon & Christiansen, 2017). The frequency of occurrence of FSs is low in general (Ellis, 2012), and thus, even highly proficient L2 learners have difficulty using FSs (Laufer & Waldman, 2011; Rafieyan, 2018). Furthermore, the focus

on vocabulary teaching or learning is mostly on words, not FSs, and there has not been enough research to obtain practical pedagogical implications (Alali & Schmitt, 2012). Hence, there is an urgent need to find an efficient way to teach and learn them. The literature review below first explains what the production effect (PE) is and how it fosters memory. Then, it argues about the applicability of PE to the area of L2 learning.

1.2 The Production Effect

Vocalization is a reasonable strategy for memorization since rehearsing is essential for holding incoming information in short-term memory (Baddeley et al., 1998). Indeed, research in psychology demonstrates that this intuitive memorization strategy is effective on the recognition and recall of items, which is called the 'production effect' (Conway & Gathercole, 1987; Fawcett & Ozubko, 2016; Gathercole & Conway, 1988; Kelly et al., 2022; MacLeod & Bodner, 2017; MacLeod et al., 2010; Ozubko, Hourihan & MacLeod, 2012). MacLeod et al. (2010) defines production effect as "the fact that producing a word aloud during study, relative to simply reading a word silently, improves explicit memory" (p. 671). The pioneering research that tested learning modalities, including vocalization, was done by Conway and Gathercole (1987). They compared various modalities in two sets of conditions: one including reading silently, mouthing silently, and vocalizing, and the other utilizing reading silently, listening, and listening and reading at the same time. Overall, the research demonstrated that conditions using spoken input resulted in better memorization. Furthermore, Gathercole and Conway (1988) compared various input modalities in five experiments, and the result showed that vocalization facilitated

the participants' memories the best. In addition, research on PE has not only exploited immediate posttests but also delayed posttests. For example, Ozubko, Hourihan, and MacLeod (2012) demonstrated that the effect of production endures one week after the learning session.

Although the effect of vocalization is robust (Conway & Gathercole, 1987; Fawcett & Ozubko, 2016; Gathercole & Conway, 1988; Kelly et al., 2022; MacLeod & Bodner, 2017; MacLeod et al., 2010; Ozubko, Hourihan & MacLeod, 2012), its theoretical explanation has yet to be agreed upon. So far, researchers have proposed three possible explanations for the production effect: a distinctiveness account (MacLeod et al., 2010), a strength account (Fawcett & Ozubko, 2016; MacLeod & Bodner, 2017), and a dual-process account (Fawcett & Ozubko, 2016). The distinctiveness account is mainly applied to the production effect in within–list designs, in which participants read items aloud or silently using one list. Hunt (2006) defines distinctiveness as "the processing of difference in the context of similarity" (p.22). This processing improves memory since distinctive items can be identified in particular events by participants at the time of testing. Applying this theory to PE, the items that are merely listened to become the background, and the items vocalized stand out in comparison to those background items, leading to better memorization of vocalized items (MacLeod et al., 2010). Previous research states that distinctiveness is enhanced by any unique process of encoding. Therefore, the greater the number of processing modes, the stronger distinctiveness becomes (Icht & Mama, 2022).

PE can also be observed in between-list designs, in which two or more lists, each of which are assigned to a different learning condition, are used. In this design, distinctiveness is absent because there

are no background items, i.e., unpronounced items. Fawcett and Ozubko (2016) explains that the reason why PE still exists in between-list designs is that vocalizing target items simply enhances their memory strength in comparison to reading them silently. This explanation is called the strength account (Fawcett & Ozubko, 2016; MacLeod & Bodner, 2017). However, the mechanism by which vocalizing items strengthens memory trace is still poorly understood. One possibility is that more attention is directed to target items when subjects vocalize than when they read silently. In fact, participants in the study conducted by Fawcett and Ozubko (2016) reported that they had paid less attention to words read silently compared to ones read aloud. The last account to be introduced is the dual-process account, which is the combination of the two explanations described above (Fawcett & Ozubko, 2016). That is, both distinctiveness and enhanced memory come into play.

1.3 PE on Second Language Learning

Although some are not conducted directly, there are studies that take into account distinctiveness and attention (Boers et al., 2012; Choi, 2017) in L2 formulaic sequence learning. Boers et al. (2012) examined whether alliteration (e.g., private property, cloth coat) in multi-word units (MWUs) produces enhanced memory. The result showed that the participants remembered alliterative MWUs better than non-alliterative MWUs. Although why alliterative MWUs were better memorized was not explained in detail, distinctiveness might have been present in the study session. In other words, non-alliterative MWUs became the background, thus making alliterative MWUs salient, resulting in the enhanced distinctiveness in alliterative MWUs. In addition to distinctiveness, attention, which is also the theoretical

base for PE, plays an important role in FSs. Choi (2017) divided 38 undergraduate students into two groups: one group using a text where collocations were typologically enhanced and the other group using unenhanced texts. The results of pretests and posttests indicated that learners with the text containing typological enhancements outperformed the other group with the text containing no enhancements. The author concluded that the better result for enhanced collocations arose since learners paid closer attention to them, which is another theoretical explanation for PE.

These facts support the potential effect of vocalization on L2 learning since the same mechanisms as those in the production effect are at work. In fact, previous research suggests that vocalizing words fosters learning of forms of nonwords (MacLeod et al., 2010), and meanings of L2 words (Ellis & Beaton, 1993; Ellis & Sinclair, 1996; Icht & Mama, 2022; Seibert, 1927). With respect to nonword forms, experiment six from MacLeod et al. (2010) investigated PE on their learning. The result demonstrated that learned nonwords were remembered better than control items, and the vocalized nonwords were better memorized than ones read silently. In light of the result, the authors concluded that preexisting representation is not a prerequisite for PE to occur, meaning PE can be applied to the learning of new word forms. Relative to studies on word form learning, PE on L2 word meanings has a long history, and a number of studies have tackled this matter (Ellis & Beaton, 1993; Ellis & Sinclair, 1996; Icht & Mama, 2022; Seibert, 1927). The streak of studies started with Seibert (1927), who showed that vocalization leads to better recall of French words than through silent reading. Ellis and Beaton (1993) compared two learning methods for words: the keyword method, which is a popular mnemonic technique for learning

foreign words, and vocalization. The result showed that vocalization was more effective for learning the meanings of German words than the keyword method. Although the studies above demonstrated that vocalization is superior to other learning conditions in terms of learning of word meaning, it was not until Icht and Mama (2022) that PE was used as an explanation for the effect of vocalization on the learning of L2 word meanings. The study showed that the words in the vocalization condition were remembered better than those in the listening condition. The author explained the result using the framework of PE. That is, distinctiveness was present, and strong memory traces were made for the vocalized items. In addition, the study examined PE on L2 vocabulary learning over the long term. The participants took a delayed posttest one or two weeks after learning L2 words. The result indicated that the effect of vocalization lasted over the two weeks.

So far, the studies that apply PE to word learning have been introduced, and the results of these studies imply that PE and its theoretical bases, distinctiveness and strong memory trace, could work in L2 learning. In fact, researchers have observed the effect of vocalization on the learning of forms (Durrant & Schmitt, 2010; Ellis & Sinclair, 1996; Ozubko, Hourihan & MacLeod, 2012) and meanings (Alali & Schmitt, 2012) of FSs as well. Ellis and Sinclair (1996) investigated whether or not vocalizing English phrases led to a better ability to recall FS forms. Participants learned ten unfamiliar Welsh words, ten phrases combining the unfamiliar Welsh words with "ble ma e___ _" ["where is___ _"], and another ten phrases consisting of the unfamiliar Welsh words and "ei___ _o" ["his___ _"]. After the learning session, the participants were instructed to orally translate English phrases into Welsh. Their speech was

recorded and whether they were able to produce the correct forms (speech) of the Welsh phrases was checked. The result demonstrated that the participants in the vocalization group produced the sounds more accurately. In addition to forms, vocalization has also been applied to the learning of the meanings of FSs (Alali & Schmitt, 2012). Alali and Schmitt (2012) investigated whether teaching L2 FSs (idioms) is different from teaching L2 single words. Tests were conducted over 12 one-hour class sessions, 6 for learning and immediate posttests and the other 6 for delayed posttests. 35 Arabic speaking intermediate school students from one intact class learned 10 single words or 10 FSs each day for a total of 60 items. In each class session, the participants learned 10 words or FSs by looking at the English forms and equivalent Arabic translations of the items. After learning, the participants took a distractive task and then proceeded to reviewing the items. In the review phase, the participants vocalized, wrote down, or did nothing, depending on the class session. The result indicated that the forms and meanings of the words and FSs vocalized were better recalled than ones not reviewed in both immediate and delayed posttests, showing a possible effect of vocalization on L2 FSs.

1.4 The Present Study

Previous research demonstrated that vocalization is more effective than just listening in the learning of forms of nonwords (MacLeod et al., 2010) and meanings of L2 words (Icht & Mama, 2022), which is also the case for the learning of forms (Ellis & Sinclair, 1996) and meanings of L2 FSs (Alali & Schmitt, 2012). Also, PE on L2 word meaning learning endures for a certain period (Icht & Mama,

2022). However, Ellis and Sinclair (1996) used the speech as a test which is in favor of the vocalization condition. Therefore, it is not clear whether the score was better because of the test type or because of the vocalization. Indeed, previous research showed that the same modalities of learning and testing foster the learning outcomes (Morris et al., 1977). Furthermore, although Alali and Schmitt (2012) demonstrated the effect of vocalization on the reviewing and learning of L2 FS meanings, whether the effect of vocalization on meanings of L2 FSs is superior to other learning modalities is unclear. Also, whether PE endures in L2 FS learning has not been tested. Thus, more validation is still needed to generalize the effectiveness of vocalization in the learning of L2 formulaic sequences. In light of the limitations above, the research questions are: (1) Does vocalization enhance form and meaning learning of formulaic sequences? and (2) If vocalization enhances L2 formulaic sequence learning, does the effect endure after a one-week interval? The present study employed verb-noun collocations as target items to obtain a sufficient number of target items (as there were many studies using verb-noun collocations) and to eliminate the influence of part of speech and the number of words. The verb-noun collocations in the present study include both verb + noun collocations and verb + article + noun collocations.

2. Method

2.1 Participants

Thirty-two undergraduate and graduate students from universities in the Tokai area (19 females, age [M = 19.78, SD = 1.83]) participated in the experiment. All participants were native Japanese

speakers majoring in various fields. They had 7.82 (*SD* = 2.82) years of English learning experience on average and had no learning, vision, or hearing disabilities. The participants' vocabulary size was measured using the Yes_No Test (Meara & Miralpeix, 2016). The result showed that the participants' vocabulary size was 4793.48 on average (*SD* = 1136.66).

2.2 Materials

Target Items

From a wide variety of collocation studies, 135 target verb-noun collocations were selected (Ackermann & Chen, 2013; Boers et al., 2017; Boers et al., 2014; Chan & Liou, 2005; Chen, 2017; El-dakhs, 2018; Eyckmans et al., 2016; Gyllstad & Wolter, 2016; Laufer & Girsai, 2008; Liu, 2010; Macis, 2018; Peters, 2014; Puimège & Peters, 2020; Snoder, 2017; Szudarski, 2012; Szudarski & Conklin, 2014; Tsai, 2020; Vilkaitem, 2016; Webb & Kagimoto, 2009; Webb et al., 2013; Wolter & Gyllstad, 2011; Wolter & Yamashita, 2014; Yamashita & Jiang, 2010; Zhang, 2017). The level of words consisting of the target verb-noun collocations were all below 3000 level in the JACET List of 8000 Basic Words (JACET Committee of Basic Words Revision, 2003). The average MI score and frequency were taken from Corpus of Contemporary American English (Table 1) (Davies, 2008). The distance between a node word and a collocate allowed four words (Durrant, 2014). All of the translations for the target verb-noun collocations were searched for in two Internet Japanese-English dictionaries: Eijiro on Web (https://www.alc.co.jp/) and Weblio (https://ejje.weblio.jp/). Since the main focus of the present

study was not the effect of frequency and MI score but the effect of vocalization, the present study did not set any criteria of frequency or MI score for target items to be verb-noun collocations.

Table 1

Information on the Target Verb-Noun Collocations

	M	*SD*	*Max*	*Min*
MI score	3.21	2.12	9.66	−1.40
Frequency	2673.12	4624.70	32671.00	26.00

Note. $N = 135$; MI = Mutual Information

Vocabulary Size Test

The Yes_No test (Meara & Miralpeix, 2016) was conducted to estimate how many English words participants knew. Participants were instructed to click "Yes" for words that they knew the meaning of and to click "Next" for words that they did not know or whose meaning they were unsure of. There were pseudowords included as well to eliminate the influence of randomly chosen answers. The test consisted of 200 items and lasted about 10 minutes. The maximum achievable score was 10,000 words, while the minimum score was zero. Ratings between 6,000 and 10,000 words are considered proficient levels. Ratings ranging from 3,500 to 6,000 words are commonly observed among intermediate-level learners. Ratings within the range of 2,000 to 3,500 words are considered beginner level.

Programming Language

The Hot Soup Processor (HSP) (Onitama, 2017) was employed to program a questionnaire, a study session, and immediate and delayed posttests.

Form and Meaning Recognition Tests

The form recognition test was based on Collex 5 (Gyllstad, 2007), in which participants were required to choose the correct combination of verbs and nouns out of three choices. Out of the three choices, one was the correct collocation that participants had learned in the study session, and the other two were combinations of verbs from other collocations and the noun for the correct answer. The frequency of occurrence for those dummy choices was adjusted to be 20 or less in COCA. On top of that, the MI scores for the dummy choices were to be 1 or less. The meaning recognition test was created based on the approach by Icht and Mama (2022). Participants were instructed to choose the best translation of verb-noun collocations from five choices, including "I don't know."

2.3 Procedure

The experiment was conducted online using Zoom, a video conferencing service. Up to 5 participants at maximum joined per video session. The programs for studying and testing, a consent form, receipts, and text files that explained the procedure of the experiment were sent to participants by e-mail before the experiment. Participants were first instructed to read the explanation for the experiment and

then asked to sign a consent form digitally. After signing the consent form, participants opened the program for the study and test sessions. Next, participants answered questions on their English learning background and proceeded to the practice session. The learning session and test session were conducted following Icht and Mama (2022). The details are shown below.

Study session. For each participant, 135 verb-noun collocations were randomly divided into three conditions (vocalization, listening, control) to eliminate item-related factors, such as corpus frequency, MI scores, and transparency of meaning. A total of 90 collocations were introduced in the study session, while the remaining 45 were presented during the test session. This meant that each participant learned a different set of collocations for each condition. In addition, the present study employed a within-subject, mixed-list design, with each participant learning 90 collocations in two conditions within one block, with the order of conditions randomized.

During the practice and study sessions, participants learned verb-noun collocations through listening or vocalization. The procedure followed the method described in Icht and Mama (2022). For both learning conditions (listening and vocalization), a written English collocation and its Japanese translation were presented for five seconds on the screen. The presentation time was longer than in Icht and Mama (2022) (four seconds vs. five seconds) because the target item in the present study was larger (word vs. formulaic sequence). An icon of an ear or mouth appeared above the English collocation (Figure 1), indicating one of the two conditions (listening or vocalization). The order in which the

conditions appeared were also randomized. After learning one collocation, a blank screen was displayed for one second. The detailed explanation for each condition is as follows.

1. Listening: The participants first saw the target collocation and heard the English collocation twice.
2. Vocalization: The participants first saw the target collocation, heard the English collocation once, and then read aloud the English collocation once.
3. Control: The collocations in the control condition did not appear in the learning session but only in the form and meaning recognition tests. The present study employed a control condition instead of asking the participants whether they knew the form and meaning of the target item to ensure that learning had occurred in the listening and vocalization conditions. This was done because some learners may guess the correct answers based on the meanings of the individual words within collocations. In such cases, learners could simply say "I knew" even if they actually had not known the meanings. Therefore, including control items and comparing the results of learned items and control items was considered a better way to determine whether learning actually occurred.

The only difference between the two learning conditions was whether or not they were vocalized, and the volume of learning was exactly the same.

Figure 1

A Screenshot of the Learning Session

[insert Figure 1.]

Test Session. After the learning session, participants took form and meaning recognition tests. Both tests included 90 verb-noun collocations that were introduced during the learning session and an additional 45 verb-noun collocations assigned to the control condition. The order of the items was randomized for both the form and meaning recognition tests. In the form recognition test, participants were given three answer choices. When they did not know the answer, they had the option to select a fourth choice, indicating "I don't know." In the meaning recognition test, participants were presented with four answer choices, and they selected a fifth choice when they were unsure of the answer. The form recognition test was administered first to prevent any influence on the meaning recognition because the meaning recognition test contained the correct forms of collocations. Following the test session, a vocabulary test was conducted. Finally, participants completed a questionnaire providing their personal

information. Approximately one week later, they took the delayed posttests, which included the same form and meaning recognition tests.

Figure 2

A Screenshot of the Form Recognition Test

[insert Figure 2.]

Figure 3

A screenshot of the Meaning Recognition Test

[insert Figure 3 here]

2.4 Statistical Analysis

Using R 4.2.1 (R core team, 2022) and the lme4 package (Bates et al., 2015), Generalized Linear Mixed-Effects Modeling (GLMM) was adopted (binomial distribution and logit link function). To avoid multicollinearity, all of the categorical variables were contrast coded. The response variables were binomially coded answers (0 for incorrect and 1 for correct answers). When any interactions were found in the models, the simple main effects were examined using the *phia* package (De Rosario-Martinez et al., 2023), and multiple comparisons were conducted using the *emmeans* (Lenth, 2023) package. Before implementing the analysis, all of the missing values were removed. There were missing values because some of the participants accidentally mistyped wrong letters in the recognition tests (e.g., 'q' instead of numbers '1', '2', '3', or '4'). The total amount of the missing values was 0.34 % of the entire data. Also, a preliminary model was fitted to confirm that learning had occurred. Items in the listening condition and the production condition were labeled as "learned" and considered as one condition. The model included the main effects of condition (learned vs. control), test types (immediate meaning vs. immediate form vs. delayed meaning vs. delayed form) and their interaction as fixed effects. The random effects were ID and items. All the models included the random slope only of test types for two reasons. First, random slopes should be included with theoretically valid reason because including all the random slopes would lead to decreased power of the model (Matuschek et al., 2017). Second, previous research (Icht et al., 2020, 2022; Icht & Mama, 2015; Swead et al., 2018) has revealed that the effect of vocalization is observable in various kinds of participants irrespective of age and disabilities such as

intellectual and hearing impairment, indicating the ubiquity of the effect of vocal production on memory. The result showed that the interaction between conditions and test types was statistically significant, and the accuracy of learned items was higher than control items in all the tests (*ps* < .001). This indicates that learning had occurred.

The following analyses were done with only the learned items, since the focus was on the different effectiveness between vocalization and listening. Also, two models were made separately for both meaning recognition tests and form recognition tests to make the analysis simple and to avoid convergence errors. Statistically insignificant variables were manually excluded from the model.

3. The Results

In general, the items in the listening condition and the vocalization condition were memorized better than those in the control condition in all the tests. Also, vocalized items were memorized better than listened ones in all the tests (Table 2).

Table 2

Descriptive Statistics of the Recognition Tests

Condition	Correct	%	Incorrect	%	Total
Immediate form recognition test (n = 32)					
Control	683	47.50	755	52.50	1438
Listening	1007	70.17	428	29.83	1435
Vocalization	1063	73.82	377	26.18	1440
Immediate meaning recognition test (n = 32)					
Control	550	38.27	887	61.73	1437
Listening	941	65.39	498	34.61	1439
Vocalization	993	69.15	443	30.85	1436
Delayed form recognition test (n = 32)					
Control	859	60.32	565	60.32	1424
Listening	1000	69.98	429	30.02	1429
Vocalization	1044	72.96	396	27.04	1431
Delayed meaning recognition test (n = 32)					
Control	557	38.82	878	61.18	1435
Listening	835	57.99	605	42.01	1440
Vocalization	859	59.74	579	40.26	1438

Note. The maximum number of trials for each condition in each test is 1440. The sums of correct and incorrect answers are different because of the 60 missing values explained in the analysis section (see section 2.4).

3.1 The Effect of Vocalization on the Learning of L2 FS Forms

GLMM was applied to investigate the effect of vocalization on L2 FS form learning. The final model for the form recognition tests ended up consisting of the learning conditions as the main effect

(Estimate = 0.178, z = 2.754, p = .006) and ID and items as random intercepts. The Akaike Information Criterion (AIC) score was 6260.800. This means that vocalized items were memorized better than listened ones (Figure 4), and the advantage was observed irrespective of the test types (immediate and delayed).

Table 1

The Final Model for the Form Recognition Tests

					Random effects	
		Fixed effects			ID	item
Parameters	Estimate	*SE*	*z*	*p*	*SD*	*SD*
Intercept	1.106	0.126	8.750	$p < .001$	0.570	0.793
Conditions: Production	0.178	0.065	2.754	$p = .006$	—	—

Note. Number of observations = 5735, n = 32. Model formula: Accuracy~ Condition + (1|ID) + (1|item).

The number of trials differed between the form recognition tests and the meaning recognition tests due to variations in the number of missing values in each of the tests.

3.2 The Effect of Vocalization on the Learning of L2 FS Meanings

The effect of vocalization on the learning of collocation meanings was also analyzed using GLMM. The Final model included conditions and test types as fixed effects and ID and item as random effects. The result demonstrated that the main effects of both the learning conditions (Estimate = 0.160, z = 2.457, p = .014) and the test types (Estimate = −0.488, z = −7.680, p < .001) were statistically significant (AIC = 6354.600). That is to say, vocalization led to a better recognition of collocation meanings, irrespective of the test types (immediate vs. delayed), and the score of the immediate meaning recognition test was higher than that of the delayed meaning recognition test (Figure 4).

Table 2

Final Model for the Meaning Recognition Tests

					Random effects	
		Fixed effects			ID	items
Parameters	Estimate	*SE*	*z*	*p*	*SD*	*SD*
Intercept	0.877	0.184	4.756	$p < .001$	0.812	1.167
Conditions: Production	0.166	0.065	2.457	$p = .014$	—	—
Test Types: Delayed Meaning	−0.488	0.064	−7.680	$p < .001$	—	—

Note. Number of observations = 5753, n = 32. Model formula: Accuracy~ Conditions + Test types + (1|ID) + (1 |item). The number of trials differed between the form recognition tests and the meaning recognition tests due to variations in the number of missing values in each of the tests.

Figure 4

The Effect of Conditions on the Form and Meaning Recognition Tests

[insert Figure 4.]

4. Discussion

4.1 The Production Effect on L2 FS Learning

In the present study, the scores of the learned items in the immediate meaning recognition test were higher than those in the delayed meaning recognition test. This means that the learning outcome of immediate meaning recognition test decayed over a week delay. On the other hand, the scores of the learned items in the immediate form recognition test were not different from those in the delayed form recognition test. This discrepancy arose because the participants had reviewed the target items in the immediate meaning recognition test, in which the correct forms of FSs were shown.

The result also revealed that vocalization is more effective than merely listening for L2 formulaic sequence (FS) learning. In addition, this effect was observed irrespective of the four test types: immediate meaning recognition, immediate form recognition, delayed meaning recognition, and delayed form recognition. This fact indicates that the effect of vocalization works on forms and meanings of FSs, and the effect endures for a week. This implies that the advantage of vocalization is purely ascribed to memory phenomena; and traits of target items such as inclusion of multi-words in target items do not affect the effectiveness of vocalization. However, the effect of vocalization was smaller compared to

previous studies on PE (e.g., Kelly et al., 2022; MacLeod et al., 2010; Zhou & MacLeod, 2021). There are two explanations for this result. First, the participants of the present study had known some of the items in the present study. Indeed, the participants scored 47.50% and 38.27% respectively for control items in the immediate form recognition test and the immediate meaning recognition test. Second, the test types in the present study are different from those in the previous PE studies. The tests employed in the previous studies were usually yes/no tests in which the participants judge whether they had learned the target items in a study session or not; thus, the measure was more sensitive to the effectiveness of vocalization.

4.2 The Theoretical Explanations of PE on L2 FS Learning

Theoretically speaking, there are three possible explanations for the result of the present study: the distinctiveness account (Hunt, 2006; MacLeod et al., 2010), the strength account (Fawcett & Ozubko, 2016; MacLeod & Bodner, 2017), and the dual-process account (Fawcett & Ozubko, 2016). According to the distinctiveness account, the more the items stand out in the background of other items, the better people can remember them (Hunt, 2006). MacLeod et al. (2010) explained that encoding distinctiveness of vocalization leads to better memorization than that of listening. That is, vocalization entailed a larger number of unique encoding processes than just listening because any uniqueness of encoding processing leads to higher distinctiveness (Icht & Mama, 2022). Icht and Mama (2022) stated that while the listening condition had only two (visual and auditory) encoding modalities, the vocalization condition had three (visual, auditory, and motor), which is the same for the present study.

The second explanation is the strength account (Fawcett & Ozubko, 2016; MacLeod & Bodner, 2017). This account states that vocalizing leaves stronger memory traces because vocalizing allows learners to pay closer attention to target items. In this study, it is also possible that the participants paid closer attention to the target FSs in the vocalization condition compared to ones in the listening condition. This account can be corroborated by the fact that vocalization leads to longer fixation time of eye-movement compared to silent reading (Rayner, 2009).

The last explanation is the dual-process account (Fawcett & Ozubko, 2016). This is a combination of the two accounts above, stating that both distinctiveness and attention are at work. This explanation seems to be the most plausible of the three explanations because both distinctiveness and attention affect L2 FS learning (Boers et al., 2012; Choi, 2017); also, PE on L2 FS form knowledge has been observed in between-list design experiments (Ellis & Sinclair, 1996), in which the theoretical basis of PE is the strength account, although different tests and procedures were employed.

5. Pedagogical Implications

The present study holds pedagogical implications based on the effectiveness of vocalization and its theoretical basis. Vocalization is an effective method for L2 learners to learn not only words but also formulaic sequences. This is because vocalization allows for effective learning even outside of the classroom, for example in a foreign language learning environment where sufficient input is not available. Thus, even simply telling learners that vocalization is effective on FSs would be helpful. As one example of the application of the present study to FS learning, learners can vocalize certain items which are especially important to them while reading other items silently in a list or a passage. In this way, vocalized FSs become distinctive and their learning will be fostered.

6. Conclusion

In the present study, it has been shown that vocal production enhances L2 formulaic sequence (FS) learning in both form and meaning recognition tests, and this effect persists over a week. The results are in line with the previous studies that have demonstrated that vocal production is effective on the learning of L2 word meaning and that the effect endures (Icht & Mama, 2022). In addition, the results indicated that vocalization not only works for learning the meaning of FSs, but also their forms.

With all the findings of the present study above, there exist limitations as well. The first limitation of the present study is that it did not include other measurements related to the production of FSs, such as recall tests and pronunciation ratings. Recall tests, in which learners orally or in writing produce a part

or a whole collocation, could have helped clarify the advantages of vocalization. This is because there is evidence that matching the learning and testing modalities enhances memorization (Morris et al., 1977). Additionally, in terms of pronunciation, it's known that a learning modality involving auditory input or vocalization can improve pronunciation (Dufour & Nguyen, 2013; Gnevsheva et al., 2021; Shao et al., 2022; Uchihara et al., 2022). The next limitation arises from the format applied to the measurement for the present study. The form and meaning recognition tests were not identical in terms of the number of choices and the format. This discrepancy arose because the present study utilized the form and meaning tests from two separate studies. In addition, the items for each condition were divided randomly for each participant. Thus, the reliability of the tests was not calculable. Those two facts on tests might have affected the results of the present study to some degree. The final limitation arises from the within-list design of the present study. The within-list design holds three possible explanations: the distinctiveness account, the strength account, and the dual-process account. On the other hand, the between-list design has only one possible explanation: the strength account. This means that using the within-list design, which of the accounts is actually at work is not clear, and whether or not vocalization enhances L2 FS learning in a between-list design still needs examining.

In order to overcome these problems, future studies should utilize the same test format, and put the same items in the same conditions. This would eliminate possibilities that the tests affect the result of future studies. Also, incorporating recall tests and pronunciation tests would allow for a broader perspective on assessing the effects of vocalization. In relation to theoretical aspects, replicates of the

present study with a between-list design should be conducted to clarify which account is appropriate for the effect of vocalization on FSs. That is, if vocalization enhances FS learning in between-list designs, whether or not the strength account actually works will be clear because it is the only explanation for the effect of vocalization in between-list designs. Alternatively, eye-movement measurement can be employed to clarify whether participants actually pay closer attention to vocalized FSs. Previous research (Ozubko, Gopie & MacLeod, 2012) holds that strong traces for vocalized items are enhanced because of heightened attention to them; thus, the use of eye-movement will also make it clear as to which of the three accounts is the most appropriate.

References

Ackermann, K., & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL) : A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235–247. https://doi.org/10.1016/j.jeap.2013.08.002

Alali, F. A., & Schmitt, N. (2012). Teaching formulaic sequences: The same as or different from teaching single words? *TESOL Journal*, *3*(2), 153–180. https://doi.org/10.1002/tesj.13

Arnon, I., & Christiansen, M. H. (2017). The role of multiword building blocks in explaining L1–L2 differences. *Topics in Cognitive Science*, *9*(3), 621–636. https://doi.org/10.1111/tops.12271

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*(1), 158–173. https://doi.org/10.1037/0033-295X.105.1.158

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4.

Journal of Statistical Software, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Boers, F., Dang, T. C. T., & Strong, B. (2017). Comparing the effectiveness of phrase-focused exercises: A partial replication of Boers, Demecheleer, Coxhead, and Webb (2014). *Language Teaching Research*, *21*(3), 362–380. https://doi.org/10.1177/1362168816651464

Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. (2014). Gauging the effects of exercises on verb-noun collocations. *Language Teaching Research*, *18*(1), 54–74. https://doi.org/10.1177/1362168813505389

Boers, F., Lindstromberg, S., & Eyckmans, J. (2012). Are alliterative word combinations comparatively easy to remember for adult learners? *RELC Journal*, *43*(1), 127–135. https://doi.org/10.1177/0033688212439997

Chan, T. P., & Liou, H. C. (2005). Effects of web-based concordancing instruction on EFL students' learning of verb-noun collocations. *Computer Assisted Language Learning*, *18*(3), 231–250. https://doi.org/10.1080/09588220500185769

Chen, Y. (2017). Dictionary use for collocation production and retention: A call-based study. *International Journal of Lexicography*, *30*(2), 225–251. https://doi.org/10.1093/ijl/ecw005

Choi, S. (2017). Processing and learning of enhanced English collocations: An eye movement study. *Language Teaching Research*, *21*(3), 403–426. https://doi.org/10.1177/1362168816653271

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45–61. https://doi.org/10.1017/S0267190512000074

Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language*, *26*(3), 341–361. https://doi.org/10.1016/0749-596X(87)90118-5

Davies, M. (2008). *The Corpus of Contemporary American English (COCA)*. https://www.english-corpora.org/coca/

Dufour, S., & Nguyen, N. (2013). How much imitation is there in a shadowing task? *Frontiers in Psychology*, *4*(346). https://doi.org/10.3389/fpsyg.2013.00346

Durrant, P. (2014). Corpus frequency and second language learners' knowledge of collocations: A meta-analysis. *International Journal of Corpus Linguistics*, *19*(4), 443–477. https://doi.org/10.1075/ijcl.19.4.01dur

Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research*, *26*(2), 163–188. https://doi.org/10.1177/0267658309349431

El-dakhs, D. A. S. (2018). What works better for collocation learning-explicit instruction or incidental learning?: A case study of Arab female undergraduate learners of English. *Electronic Journal of Foreign Language Teaching*, *15*(1), 39–54. https://e-flt.nus.edu.sg/wp-content/uploads/2020/09/el-dakhs.pdf

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17–44. https://doi.org/10.1017/S0267190512000025

Ellis, N. C., & Beaton, A. (1993). Factors affecting the learning of foreign language vocabulary: Imagery keyword mediators and phonological short-term memory. *The Quarterly Journal of*

Experimental Psychology Section A, *46*(3), 533–558. https://doi.org/10.1080/14640749308401062

Ellis, N. C., & Sinclair, J. (1996). Working memory in the acquisition of vocabulary and syntax: Putting language in good order. *The Quarterly Journal of Experimental Psychology*, *49*(1), 234–250. https://doi.org/10.1080/713755604

Eyckmans, J., Boers, F., & Lindstromberg, S. (2016). The impact of imposing processing strategies on L2 learners' deliberate study of lexical phrases. *System*, *56*, 127–139. https://doi.org/10.1016/j.system.2015.12.001

Fawcett, J. M., & Ozubko, J. D. (2016). Familiarity, but not recollection, supports the between-subject production effect in recognition memory. *Canadian Journal of Experimental Psychology = Revue Canadienne de Psychologie Experimentale*, *70*(2), 99–115. https://doi.org/10.1037/cep0000089

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). http://tinyurl.com/carbook

Gathercole, S. E., & Conway, M. A. (1988). Exploring long-term modality effects: vocalization leads to best retention. *Memory & Cognition*, *16*(2), 110–119. https://doi.org/10.3758/bf03213478

Gnevsheva, K., Szakay, A., & Jansen, S. (2021). Phonetic convergence across dialect boundaries in first and second language speakers. *Journal of Phonetics*, *89*, 101110. https://doi.org/10.1016/j.wocn.2021.101110

Gyllstad, H. (2007). *Testing English collocations: Developing receptive tests for use with advanced Swedish learners*. Lund University Press. https://lucris.lub.lu.se/ws/files/5893676/2172422.pdf

Gyllstad, H., & Wolter, B. (2016). Collocational processing in light of the phraseological continuum model: Does semantic transparency matter? *Language Learning*, *66*(2), 296–323. https://doi.org/10.1111/lang.12143

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In *Distinctiveness and memory* (pp. 3–25). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195169669.003.0001

Icht, M., Ben-David, N., & Mama, Y. (2020). Using vocal production to improve long-term verbal memory in Adults with intellectual disability. *Behavior Modification*. https://doi.org/10.1177/0145445520906583

Icht, M., & Mama, Y. (2015). The production effect in memory: A prominent mnemonic in children. *Journal of Child Language*, *42*(5), 1102–1124. https://doi.org/10.1017/S0305000914000713

Icht, M., & Mama, Y. (2022). The effect of vocal production on vocabulary learning in a second language. *Language Teaching Research*, *26*(1), 79–98. https://doi.org/10.1177/1362168819883894

Icht, M., Taitelbaum-Swead, R., & Mama, Y. (2022). Production improves visual and auditory text memory in younger and older adults. *Gerontology*, *68*(5), 578–586. https://doi.org/10.1159/000518894

JACET Committee of Basic Words Revision. (2003). *JACET list of 8000 basic words: JACET 8000*.

Japan Association of College English Teachers. https://ci.nii.ac.jp/ncid/BB21043002

Kelly, M. O., Ensor, T. M., Lu, X., MacLeod, C. M., & Risko, E. F. (2022). Reducing retrieval time modulates the production effect: Empirical evidence and computational accounts. *Journal of Memory and Language*, *123*, 104299. https://doi.org/10.1016/j.jml.2021.104299

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics*, *29*(4), 694–716. https://doi.org/10.1093/applin/amn018

Laufer, B., & Waldman, T. (2011). Verb-Noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*, 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Liu, D. (2010). Going beyond patterns: Involving cognitive analysis in the learning of collocations. *TESOL Quarterly*, *44*(1), 4–30. https://doi.org/10.5054/tq.2010.214046

Macis, M. (2018). Incidental learning of duplex collocations from reading : Three case studies. *Reading in a Foregn Language*, *30*(1), 48–75. https://doi.org/10125/66738

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science*, *26*(4), 390–395. https://doi.org/10.1177/0963721417691356

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: delineation of a phenomenon. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*(3), 671–685. https://doi.org/10.1037/a0018785

Matuschek, H., Kliegl, R., Vasishtha, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Meara, P., & Miralpeix, I. (2016). *Tools for researching vocabulary*. Multilingual Matters. https://doi.org/10.21832/9781783096473

Morris C, D., Bransford, J. D., & Franks J, J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Onitama. (2017). *Hot Soup Processor (Version 3.5)*. http://hsp.tv/

Ozubko, J. D., Gopie, N., & MacLeod, C. M. (2012). Production benefits both recollection and familiarity. *Memory and Cognition*, *40*(3), 326–338. https://doi.org/10.3758/s13421-011-0165-1

Ozubko, J. D., Hourihan, K. L., & MacLeod, C. M. (2012). Production benefits learning: The production effect endures and improves memory for text. *Memory* , *20*(7), 717–727. https://doi.org/10.1080/09658211.2012.699070

Peters, E. (2014). The effects of repetition and time of post-test administration on EFL learners' form recall of single words and collocations. *Language Teaching Research*, *18*(1), 75–94. https://doi.org/10.1177/1362168813505384

Puimège, E., & Peters, E. (2020). Learning formulaic sequences through viewing L2 television and

factors that affect Learning. *Studies in Second Language Acquisition*, *42*(3), 525–549. https://doi.org/10.1017/S027226311900055X

R core team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rafieyan, V. (2018). Role of knowledge of formulaic sequences in language proficiency: significance and ideal method of instruction. *Asian-Pacific Journal of Second and Foreign Language Education*, *3*(1), 1–23. https://doi.org/10.1186/s40862-018-0050-6

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506. https://doi.org/10.1080/17470210902816461

Seibert, L. C. (1927). An experiment in learning French. *Journal of Educational Psychology*, *18*(5), 294–309. https://doi.org/10.1037/h0074206

Shao, Y., Saito, K., & Tierney, A. (2022). How does having a good ear promote instructed second language pronunciation development? Roles of domain‐general auditory processing in choral repetition training. *TESOL Quarterly*. https://doi.org/10.1002/tesq.3120

Snoder, P. (2017). Improving English learners' productive collocation knowledge: The effects of involvement load, spacing, and intentionality. *TESL Canada Journal*, *34*(3), 140–164. https://doi.org/ttps://doi.org/10.18806/tesl.v34i3.1277

Swead, R. T., Mama, Y., & Icht, M. (2018). The effect of presentation mode and production type on

word memory for hearing impaired signers. *Journal of the American Academy of Audiology*, *29*(10), 875–884. https://doi.org/10.3766/jaaa.17030

Szudarski, P. (2012). Effects of meaning-and form-focused instruction on the acquisition of verb-noun collocations in L2 English. *Journal of Second Language Teaching & Research*, *1*(2), 3–37. https://pops.uclan.ac.uk/index.php/jsltr/article/view/32/15

Szudarski, P., & Conklin, K. (2014). Short- and long-term effects of rote rehearsal on ESL learners' processing of L2 collocations. *TESOL Quarterly*, *48*(4), 833–842. https://doi.org/10.4135/9781452276274.n840

Tsai, M.-H. H. (2020). The effects of explicit instruction on L2 learners' acquisition of verb–noun collocations. *Language Teaching Research*, *24*(2), 138–162. https://doi.org/10.1177/1362168818795188

Uchihara, T., Webb, S., Saito, K., & Trofimovich, P. (2022). Does mode of input affect how second language learners create form–meaning connections and pronounce second language words? *Modern Language Journal*, *106*(2), 351–370. https://doi.org/10.1111/modl.12775

Vilkaitem, L. (2016). Are nonadjacent collocations processed faster? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *42*(10), 1632–1642. https://doi.org/10.1037/xlm0000259

Webb, S., & Kagimoto, E. (2009). The effects of vocabulary learning on collocation and meaning. *TESOL Quarterly*, *43*(1), 55–77. https://doi.org/10.1002/j.1545-7249.2009.tb00227.x

Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning*, *63*(1), 91–120. https://doi.org/10.1111/j.1467-9922.2012.00729.x

Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics*, *32*(4), 430–449. https://doi.org/10.1093/applin/amr011

Wolter, B., & Yamashita, J. (2014). Processing collocations in a second language: A case of first language activation? *Applied Psycholinguistics*, *36*(5), 1193–1221. https://doi.org/10.1017/S0142716414000113

Wray, A., & Perkins, M. (2000). The functions of formulaic language: An integrated model. *Language and Communication*, *20*(1), 1–28. https://doi.org/10.1016/S0271-5309(99)00015-4

Yamashita, J., & Jiang, N. (2010). L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Quarterly*, *44*(4), 647–668. https://doi.org/10.5054/tq.2010.235998

Zhang, X. (2017). Effects of receptive-productive integration tasks and prior knowledge of component words on L2 collocation development. *System*, *66*, 156–167. https://doi.org/10.1016/j.system.2017.03.019

Zhou, Y., & MacLeod, C. M. (2021). Production between and within: distinctiveness and the relative magnitude of the production effect. *Memory* , *29*(2), 168–179. https://doi.org/10.1080/09658211.2020.1868526

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45–61.

De Rosario-Martinez, H., Fox, J., & R Core Team. (2023). *Package "phia."* https://github.com/heliosdrm/phiahttps://cran.r-project.org/web/packages/phia/index.html

Lenth, R. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.4.8*. https://cran.r-project.org/package=emmeans