International Journal of English Linguistics; Vol. 11, No. 5; 2021
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

The Effects of Two Association Measures on L2 Collocation Processing

Alaa Alzahrani[1]

[1] Department of English Language & Literature, King Saud University, Riyadh, Saudi Arabia

Correspondence: Alaa Alzahrani, 3135 Mihyar Ad Daylami, Riyadh, Saudi Arabia. E-mail: alaa-zahrani@hotmail.com

Received: May 8, 2021 Accepted: August 7, 2021 Online Published: August 18, 2021

doi:10.5539/ijel.v11n5p28 URL: https://doi.org/10.5539/ijel.v11n5p28

Abstract

The influence of association measures has been little examined in research on L2 collocation processing. For this reason, the present study replicated Öksüz et al. (2020) experiment on intermediate L2 learners of English to determine whether the association measure mutual information (MI) is a stronger predictor of L2 performance than the Log Dice measure. Twenty-two intermediate Arab learners of English completed a timed acceptability judgment task on the online Gorilla platform. The task included (1) high-frequent collocations (e.g., bad news), (2) low-frequent collocations (e.g., only friend), and (3) non-collocates (e.g., true news, wrong friend) which had differing MI and Log Dice scores. Mixed-effects models were built to analyze the participants' reaction times to the three conditions. The results showed that the frequency of the collocation (operationalized as item type) and its length significantly influenced reaction times, while both MI and Log Dice scores did not surface as significant predictors. This suggests that intermediate English L2 learners are not sensitive to corpus-based association measures. The results have important implications for L2 teaching and testing and may indicate that it is not worthwhile to determine which collocations to include in the materials based mainly on the strength of the association.

Keywords: association measures, collocations, L2 learners, MI scores, Log Dice scores

1. Introduction

Collocations have received considerable attention in Second Language Acquisition (SLA) research as they are essential for fluency for second language (L2) learners (Henriksen, 2013; Wray, 2012). Collocations are defined as "frequently recurring two-to-three-word syntagmatic units which can include both lexical and grammatical words, e.g., verb + noun (pay tribute), adjective + noun (hot spice), preposition + noun (on guard) and adjective + preposition (immune to)" (Henriksen, 2013, p. 30). The importance of collocations in SLA research stems partly from the finding that skillful use of collocations predicts the quality of L2 writing (Bestgen, 2017; Bestgen & Granger, 2014) and L2 oral proficiency (Xu, 2018). This suggests that acceptable use of collocations is a predictor of L2 learners' proficiency level. Due to the facilitative role of collocation in L2 proficiency, researchers have compiled collocations lists (Ackermann & Chen, 2013; Durrant, 2009) and developed a web-based collocation assistant (Frankenberg-Garcia et al., 2019) to help L2 learners in producing the most appropriate collocation sequences. Despite these efforts, it has been constantly reported that L2 learners find it challenging to acquire and appropriately use collocations (Laufer & Waldman, 2011; Nesselhauf, 2005; Nguyen & Webb, 2017). This limited ability to acquire the appropriate native-like knowledge of collocations puts the L2 learner at a disadvantage. Knowledge of formulaic sequences such as collocations helps L2 and native speakers process language faster and communicate better (Wray, 2012).

In the SLA literature, collocations are commonly identified using a large corpus and specific association measures (Henriksen, 2013). A corpus is a collection of texts, whether spoken or written, stored and accessed electronically (McEnery & Hardie, 2011). Examples of large corpora commonly used in collocation research include British National Corpus (BNC) (Davies, 2004) and Corpus of Contemporary American English (COCA) (Davies, 2008). The association measure assesses how strongly two words such as "ideal" and "solution" are associated. The strength of the relationship of the word pair "ideal solution" is calculated by comparing between the number of its occurrence as a word combination "ideal solution" and the individual words' occurrence with other candidate words, e.g., ideal way, ideal cup, ideal friend, different solution, new solution, best solution. If the word pair "ideal solution" appears together more frequently in the corpus than appearing individually, the association score for the word pair would be high as the relationship between the word pair is strong. Word pairs with high association scores such as "ideal solution" are identified as a collocation.

Different association measures help in highlighting collocations in a large corpus (Evert, 2008). The most common association measure used in SLA and psycholinguistic research is Mutual Information (MI) score (e.g., Ellis et al., 2008; Sonbul, 2015; Yi, 2018). The MI measure gives high scores for word pairs that almost exclusively co-occur together and low scores for those that do not. One limitation of the MI measure is that it gives higher scores to word pairs that rarely occur in a corpus (Brezina, 2018). In other words, the MI measure has a bias for low frequent word combinations as it gives them higher scores compared to high-frequent combinations (see the MI and Log Dice section for examples). For this reason, the reliance on MI in SLA and psycholinguistic research has been criticized as it might identify rare collocations which are not familiar to L2 learners (Gablasova et al., 2017). Testing L2 learners on collocations they may not know and thus have not acquired yet to determine their knowledge of collocations is invalid. Therefore, another association metric known as Log Dice has been suggested to be a better predictor of L2 learners' collocation knowledge because it does not have the disadvantage of the MI measure (Gablasova et al., 2017). A description of both Log Dice and MI is provided in the next section.

2. Theoretical Background and Previous Research

2.1 Identifying Collocation

Two main approaches are used in identifying collocations in the literature: a phraseological approach (Cowie, 1998) and a frequency-based one (Sinclair, 1991). While the phraseological method decides whether a word pair such as "underlying cause" constitutes a collocation or not by relying on the grammatical/semantic properties of the individual words, i.e., "underlying" and "cause," the frequency-based approach uses certain corpus-derived association statistics to do so. The latter approach is more adopted in collocation processing research because corpus statistics (raw frequency, MI score, Log Dice score) examine frequency effects on collocation processing.

As the frequency-based approach in identifying collocations is more widespread in the processing literature, the main theoretical framework adopted in this area is the usage-based model, which emphasizes the role of frequency. The usage-based model explains how formulaic language, including collocations, is stored and processed (Bybee, 1998, 2006; Tomasello, 2003). This model proposes that collocations (e.g., last night) are processed faster (i.e., read and recognized faster) than non-collocates (e.g., violent night) because language users encounter collocations many times in native speech and eventually become sensitive to their frequency. For this model, then, frequent exposure makes people recognize collocations faster. Thus, frequency is an important factor that explains how collocations are recognized quicker than non-collocations.

2.2 Defining MI and Log Dice

Although collocations can be identified using different corpus metrics (Evert, 2008; Rychly, 2008), only three metrics are reviewed here as they are examined in this study. These metrics are (a) raw frequency, (2) MI score (MI), and (3) Log Dice score. The Raw frequency, sometimes called absolute frequency, calculates only the total number of occurrences of the collocation in a specific corpus (Brezina, 2018). For example, the collocation "underlying cause" has occurred 395 times in COCA, and thus its raw frequency is 395. Nevertheless, the raw frequency only gives information about how many times a word pair occurred in a corpus; it does not distinguish between frequent collocations "underlying issues" (freq. = 188) and frequent non-collocates "its underlying" (freq. = 168).

To address this issue, other metrics were used along with raw frequency to identify collocations. These metrics are labeled as association metrics as they (a) calculate the frequency of the word pair and (b) measure the strength of the relationship between a word pair, i.e., between "underlying" and "issues". To better understand how association measures work, it is important to consider (a) their statistical aim and features, (b) the resulting values and their interpretation, and (c) the main characteristics of the collocations they highlight. In the remaining part of this section, these three aspects are discussed and compared for the two examined association measures in this study: the MI score and the Log Dice score.

Both MI and Log Dice are measures of effect size. They aim to answer the question "how strongly are the words attracted to each other?" (Evert, 2008, p. 1228). Also, both measures can be used to compare collocations extracted from corpora of different sizes. In other words, size differences between data samples are accounted for as the two metrics use a method to correct for size variation. Although MI score and Log Dice score share some statistical features, they do not highlight the same list of collocations. This is because the MI differs from Log

Dice in some respects, which are explained next.

While MI scores can range from 1 to 30 and beyond, the Log Dice scores have a more fixed value ranging from 1 to 14 (Rychly, 2008). Nevertheless, higher MI and Log Dice scores are seen as an indication of stronger association (Evert, 2008). However, unlike the Log Dice score, the MI score is prone to a low-frequency bias. That is, the MI tends to identify strongly associated collocations that are rarely used, such as proper names and specialized terms (e.g., carbonic anhydrase). In fact, Gablasova et al. (2017) observed that the MI score is sometimes higher for lower-frequency collocations compared to higher-frequency ones. For instance, while the word pair "ceteris paribus" has a higher frequency (freq. = 46) than "jampa ngodrup" (freq. = 10), it has a lower MI score (MI = 21) than the latter (MI = 23.2). This tendency to identify low-frequency word pairs makes the MI score fail to spot collocations that are equally distributed across various linguistic situations (e.g., conversations, fiction, news, academic contexts) (Gablasova et al., 2017). Thus, MI is limited in two aspects. It rewards low-frequent word pairs with higher scores and highlights specialized terms and names as strong collocations (McEnery & Hardie, 2011). Not surprisingly, studies have reported that L2 learners could not recognize collocations with high MI scores, suggesting its inappropriateness in gauging L2 learners' collocation knowledge (Gablasova et al., 2017).

These problems of the MI score are not present in the Log Dice score. The reason is that the Log Dice does not favor low-frequency collocations; both low-frequent word pairs and high-frequent ones receive similar Log Dice scores if they are strongly associated with each other (i.e., almost always appear together) (Brezina, 2018). This indicates that the collocations with high Log Dice scores are more likely to be recognized by L2 learners. Such collocations are more general and occur in various contexts rather than being restricted to specialized contexts (e.g., finance, science, politics). Thus, the Log Dice score highlights widely dispersed collocations, i.e., used commonly in different fields. It has been noted that the dispersion of collocations predicts L2 learners' acquisition of collocations because the degree of dispersion informs us about how widely the collocation is used (Gablasova et al., 2017). In other words, collocations with high Log Dice scores are more likely to be recognized by L2 learners and, as such, are more appropriate for testing their collocation knowledge. Another advantage of the Log Dice score is that it is easier to interpret compared to the MI score as its values do not exceed 14, as was mentioned above (Rychly, 2008). Despite these strengths of the Log Dice score, SLA and psycholinguistic research have largely relied on MI score and t-score (Bestgen & Granger, 2014; Ellis et al., 2008; Sonbul, 2015) with limited interest in the Log Dice score.

2.3 L2 Processing Studies on Association Measures

Several studies have looked at the impact of association measures on the processing of formulaic sequences by L2 learners of English. Two types of formulaic sequences were mostly examined, including lexical bundles, e.g., "at the beginning of", "I don't want that" (Ellis et al., 2008; McCauley & Christiansen, 2017) and collocations (González Fernández & Schmitt, 2015; Öksüz et al., 2020; Yi, 2018). While the lexical bundle studies reported similar findings, contradictory results were reported in collocation studies. For instance, the works examining the effects of association measures on lexical bundles have found that L2 speakers are not sensitive to MI score as their performance in the tasks was not explained by MI score. On the other hand, some studies of collocation processing reported that L2 learners' responses are not influenced by the MI score (González Fernández & Schmitt, 2015), influenced to some extent (Yi, 2018) or strongly influenced (Öksüz et al., 2020). Further, while most L2 processing studies examined the effects of MI score, only one study investigated the Log Dice score (Öksüz et al., 2020), which calls for more research on this new association measure to understand its effects on processing.

One main study which investigated the effects of MI score on L2 learners' processing of formulaic language was done by Ellis et al. (2008). In this work, Ellis et al. conducted three experiments using the same items across the three experiments but recruited different participants for each experiment. In experiment 1, a grammaticality judgment task was used to assess how English natives and 11 English as a Second Language (ESL) learners recognize 108 formulaic sequences with differing a) MI scores and b) raw frequency of occurrence. The ESL participants were graduate students from different language backgrounds (Chinese, Thai, Korean and Spanish). However, their proficiency level was not reported but rather hinted at. The authors labeled their ESL participants as "advanced learners" and mentioned that their "English language proficiency was sufficient to permit enrollment at the university for a graduate degree through the medium of English" (p. 383). The results of experiment 1 showed that while native speakers were more sensitive to sequences with high MI scores, their ESL counterparts were influenced by the raw frequency of occurrence, not the MI. Similar findings were found in experiment 2 that examined the production of formulaic sequences in a read-aloud method. However, in experiment 3 using a priming production task, the effects of raw frequency of occurrence on ESL participants' performance were not present. A forced entry multiple regression analysis was carried out for the results of experiment 3. It showed a relatively substantial effect of MI at = -0.20 on ESL learners' production of formulaic sequences and a stronger one for natives (β = -0.47). To summarize, while the MI score predicted English natives processing time in both recognition (experiment 1) and production (experiment 2 and 3) tasks, it did not determine ESL learners' processing in two receptive and productive tasks (experiment 1 and 2). It only showed limited effects in the last primed productive task (experiment 3).

Similar to Ellis et al. (2008), it was reported in a simulation study by McCauley and Christiansen (2017) that adult L2 speakers' production of lexical bundles is explained by the times they encountered them (raw frequency) rather than the lexical bundles' MI score. However, the findings of both studies are limited as they included a small number of L2 participants. The number of the L2 participants in Ellis et al. (2008) was quite small, ranging from 6 to 16, limiting the extent of generalization of the results. Likewise, the L2 data in McCauley and Christiansen (2017) were based on recorded interactions of 7 learners.

Other studies considered the effects of association measures on L2 learners' processing of collocations. One is a study by González Fernández and Schmitt (2015), who investigated 108 Spanish learners of English to measure their productive knowledge of 50 collocations in a productive form recall test. The 50 collocations had different a) corpus frequency, b) t-score, and c) MI score because the researchers wanted to determine which collocation identification method best relates to the collocation knowledge of the examined learners. Another instrument used in the study was a questionnaire to collect information about how participants use L2 and in which activities. The 108 participants were divided into three groups: high, mid, and low proficiency based on their self-rating of their L2 proficiency. The analysis revealed that knowledge of the 50 collocations correlated moderately with corpus frequency (r = .45) and the daily use of English outside the classroom in activities such as reading, watching movies, and social networking (r = .56). Nevertheless, the researchers concluded that corpus frequency is still not a strong factor, as it only explained around 20% of the collocation knowledge examined in the test. The limited effects of MI score in this study might be since González Fernández and Schmitt (2015) examined learners from different proficiency levels, which was not the case in the following two studies.

The results found in González Fernández and Schmitt's (2015) study were not confirmed in more recent works. For example, Yi (2018) examined several factors, including the effect of MI on L1 and advanced L2 speakers' processing of 180 adj-noun collocations (e.g., old age, real world). The proficiency level of the Chinese L2 learners of English was reported as advanced based on (a) their most recent TOEFL iBT score and (b) cloze test score. An acceptability judgment task was administrated using DMDX software. The participants' reaction time and accuracy were recorded and analyzed using the lme4 package in R. The reaction time analysis showed that the MI score affected L2 speakers' processing speed more than L1 speakers. However, the accuracy analysis showed that an increase in MI scores improved only L1 speakers' chance of making a correct response but not their L2 counterparts.

Although Yi (2018) concluded that advanced L2 speakers are more influenced by the MI score than L1 speakers, this conclusion does not seem to be supported by the results of one of the analyses reported in the study. The accuracy analysis done in Yi (2018) revealed that collocations with high MI scores might pose some problems for L2 learners. That is, L2 learners might respond quickly to high MI-collocations (in less than a second), but their response is likely to be incorrect. This suggests that assessing advanced L2 learners' receptive knowledge of collocations based on the MI score of the collocations might not be valid as those learners may not have fully integrated their meaning.

Another study that confirmed the influence of MI score on L2 collocation is by Öksüz et al. (2020). In this study, Öksüz et al. (2020) examined the effects of 1) individual word frequency, 2) phrase frequency, and 3) association measures (MI and Log Dice) on L1 and L2 processing of 120 adj-noun collocations in an acceptability judgment task. The study had two main objectives. One was to compare between the influence of single-word frequency (e.g., long, time) and phrase frequency (e.g., long time) on the recognition of highly frequent collocations (e.g., long time) and low-frequent ones (e.g., similar case). The study's second goal was to examine whether L1 and L2 speakers recognize collocations based on their association measure (i.e., how strong is the relationship between two words and whether they always appear together). For these reasons, the study included collocations that varied in frequency, MI score, and Log Dice score. The items were divided into (a) high frequency collocations (frequency = ≥ 300, LogDice = ≥ 7), (b) low-frequency ones (frequency = 10 to 150, LogDice = 2 to 4), and (c) control items as a baseline for comparison (frequency ≤ 10, LogDice = -0.93). The control adj-noun items consisted of random combinations of the nouns used for the main items and new adjectives that were not used for the main items (e.g., dirty time VS long time). Thus, the participants saw each noun twice in different conditions, making them process the nouns quickly because they were previously exposed to them. To solve this problem, the authors had the items presented in a randomized order.

Thirty English native speakers and 32 advanced Turkish learners of English studying in Turkey participated in the study. The proficiency level of the L2 participants was measured based on their scores in a vocabulary test called LexTALE (M = 84.85) and their self-rating of their perceived proficiency in the four skills: writing, reading, speaking, and listening. It was noted that 21 of the L2 participants had lived in an English-speaking country for more than one month. The acceptability judgment task was administered using the PsychoPy software. The task required the participants to indicate whether they thought the 120 items are acceptable combinations in English or not. A practice session was completed before doing the acceptability judgment task to familiarize the participants with the task.

The analysis of participants' responses to the timed grammaticality judgment task was as follows. The analysis was run using the lme4 package on R, and mixed-effects models were constructed to compare response times between the participating groups and item conditions. Only the items that received a correct response were included in the mixed-effect models. It was found that both the single-word frequency and phrase frequency affected L1 and L2 speakers' speed of responding in the task, with a reduced effect of single-word frequency for the high-frequency collocations. It was also reported that the association measures (MI and Log Dice) had similar effects on the processing of collocations for both L1 and L2 speakers. That is, both groups were equally sensitive to collocations with high MI and Log Dice scores as they processed them faster than those with lower MI and Log Dice scores. These findings suggest that L1 and advanced L2 speakers process collocations in the same way. While Öksüz et al. (2020) examined MI in combination with the new association measure Log Dice, its findings might apply to advanced L2 learners only as other proficiency levels were not investigated. The effect of proficiency level on the processing of collocations has been observed in previous works with a processing advantage for more advanced levels compared to lower ones (Sonbul, 2015).

Having reviewed the related studies, it can be observed that studies examining L2 collocation processing reported different findings of the effects of association measures. Unlike Öksüz et al. (2020), which found a strong effect for both MI and Log Dice, González Fernández and Schmitt (2015) and Yi (2018) revealed varying degrees of MI score effect. Two possible reasons can explain this difference in results. One is related to the examined L2 learners' proficiency, and the other is related to the type of analysis carried in the study. It is possible that limited influence of MI score was observed in González Fernández and Schmitt (2015) because they included learners at three different L2 stages, while only advanced learners were examined in both Öksüz et al. (2020) and Yi (2018). Another possible explanation for the conflicting results is that whereas Öksüz et al. (2020) focused on analyzing the reaction time (the time it took the L2 participant to respond), González Fernández and Schmitt (2015) and Yi (2018) analyzed the accuracy of L2 participants' responses (whether they correctly identified collocations and rejected non-collocates). These points of difference are outlined in Table 1 below.

Table 1. A summary of the reviewed studies on L2 collocation processing

Study	Examined L2 Proficiency	Task	Effect of MI	Effect of Log Dice
González Fernández and Schmitt (2015)	Advanced Intermediate Beginner	Untimed Recall test	Limited effect (accuracy analysis)	
Yi (2018)	Advanced	Timed acceptability judgment task	Limited effect (accuracy analysis) Strong effect (reaction analysis)	
Öksüz et al. (2020)	Advanced	Timed acceptability judgment task	Strong effect (reaction analysis)	Strong effect (reaction analysis)

In addition to the conflicting evidence on the effects of MI score, the review reveals that only one study examined the Log Dice as a predictor of L2 processing of collocations. Most L1 and L2 studies investigating the processing of different types of formulaic language have solely considered the influence of MI score (Ellis et al., 2008; González Fernández & Schmitt, 2015; McCauley & Christiansen, 2017; Yi, 2018), with limited attention to the Log Dice score (Öksüz et al., 2020). It was discussed in the MI and Log Dice section that Log Dice does not have the limitation of MI (a bias towards collocations with a low frequency). As such, it was argued that collocations with high Log Dice scores are more likely to be familiar to L2 learners, unlike collocations with high MI scores. This is because the Log Dice measure identifies general collocations that are used across various contexts rather than those which are highly specialized. Despite the availability of a more appropriate association measure, Log

Dice has been little examined in research on L2 collocation processing. Öksüz et al. (2020) considered the effects of Log Dice, but it only examined advanced L2 learners' (as measured by their scores on a vocabulary test), and the results might not apply to L2 learners at lower proficiency levels. Thus, the present study aims to replicate Öksüz et al.'s (2020) experiment on intermediate L2 learners of English to determine whether one association measure is a stronger predictor of L2 performance. Specifically, this study will compare between (a) MI scores and (b) Log Dice scores in predicting intermediate Arab L2 learners' acceptance of high-frequent collocations (e.g., bad news), low-frequent collocations (e.g., only friend), and non-collocates (e.g., true news, wrong friend) in a timed acceptability judgement task.

Determining which association measure is a better predictor of L2 learners' processing of collocation has implications for language learning and testing. If learners are more familiar with collocations with high Log Dice scores, we can encourage L2 language teachers and material designers to focus on teaching, and testing collocations with high Log Dice scores rather than those with a high MI score only. The significance of examining association measures has been noted in the literature as these measures affect research findings and the research-based insights into language learning and processing (Gablasova et al., 2017; Gries, 2013). With this in mind, the present study is set out to address the following research question: Are intermediate-level English L2 speakers sensitive to collocations identified by MI more than those identified by Log Dice scores?

3. Method

3.1 Participants

This study recruited 22 female Saudi L2 learners of English. Due to gender segregation in Saudi higher education, only female students from one university were recruited because they have relatively similar L2 instructional backgrounds. The participants were in their final or prefinal year in an English language undergraduate program because, at this stage, they are more likely to have reached the intermediate level. The present study examined intermediate English learners as most previous research focused on advanced learners (Öksüz et al., 2020; Yi, 2018). Purposeful sampling was used to recruit intermediate female Arab learners of English since this sampling technique would ensure the selection of participants with the required proficiency level.

As shown in Table 2, most participants were in the 20–24 age group (91%) going through the final or pre-final year of undergraduate English studies. Only two lived in an English-speaking country; one lived for four years and the other for five months. Nevertheless, their scores on the LexTALE test were similar to those who did not go to an English-speaking country, suggesting limited effects of their stay abroad on their vocabulary knowledge. For this reason, the data of the participants who lived abroad were included in the analysis. Also, most of the participants started learning English before they turned 15, with only four of them learning after the age of 15.

Table 1. The participants' background information in frequencies and percentages

Characteristics	Age group		Year of study		Lived abroad		Learned English	
	20–24	25–29	Final	Pre-final	Did not	Did	Before 15	After 15
Frequency (%)	20 (91)	2 (9)	17 (77)	5 (23)	20 (91)	2 (9)	18 (82)	4 (18)

Following Öksüz et al. (2020), the participants' language proficiency was measured by their performance on the free web-based LexTALE test to allow more accurate comparison between the two studies. This test is designed to assess intermediate to advanced L2 learners' written receptive vocabulary knowledge in English using yes/no questions (Lemhöfer & Broersma, 2012). This proficiency measure has been found to be correlated with a general English proficiency measure (TOEIC) and a good predictor of English vocabulary knowledge (Lemhöfer & Broersma, 2012).

The recruited participants formed a relatively homogenous group from the intermediate proficiency level who differed from the advanced L2 participants in Öksüz et al. (2020). To compare the two groups' average LexTALE scores, the parametric independent sample t-test was run since the data set meets the assumption of homogeneity of variances (F = 42.8, p = .713). Also, bootstrapped CI was used and reported to avoid any potential violation of normal distribution. An independent sample t-test revealed a significant difference in the average LexTALE scores between the intermediate L2 participants in this study (N = 22, *M* = 52.6, *SD* = 1.2, BCa 95% CI [54.5, 58.7]) and the advanced L2 group in Öksüz et al.'s (2020) (N = 32, *M* = 84.8, *SD* = .92, BCa 95% CI = [83, 86.7]), *t* (52) = 18.7, p < .001, BCa 95% CI = [25, 31]).

3.2 Instruments

This study followed the tasks used in Öksüz et al. (2020) and adopted with no changes the items of the acceptability judgment task from Öksüz et al. First; the study used the free web-based LexTALE test to assess the participants' English vocabulary knowledge using Yes/No format. This measure was chosen to guarantee a more accurate comparison between the participants' performance in the present study and in Öksüz et al.'s (2020) study. Second, an acceptability judgment task was created and conducted using Gorilla Experiment Builder's online platform (Anwyl-Irvine et al., 2020; http://www.gorilla.sc) to abide by COVID-19 safety measures. This platform was specifically chosen because it can reliably run reaction-time-sensitive experiments (Anwyl-Irvine et al., 2020). Thirty high-frequency adjective-noun collocations (e.g., long time), 30 low-frequency ones (e.g., tiny room), and 60 non-collocate controls (e.g., dirty time, key room) were presented to the participants. This creates three-item conditions: (a) high-frequency collocations, (b) low-frequency collocations, and (b) baseline non-collocates. Thus, a total of 120 items were included in the task. Examples from the three item conditions are included in Table 3.

Table 2. Examples of the target and control word pairs extracted from the BNC XML corpus (adopted from Öksüz et al., 2020)

Condition	Word pair	Pair frequency (per million words)	MI score	Log Dice score
High frequency	Long time	4.61	3.15	7.28
	Young people	4.53	4	7.36
Low frequency	Inner world	2.85	2.39	2.92
	Difficult life	3.11	1.01	3.76
Control	Dirty time	1.41	-3.03	-3.22
	Sudden people	1.54	-2.92	-2.48
	Necessary world	1.64	-3.64	-1.14
	Final life	1.9	-2.51	-0.22

The acceptability judgment task required the participants to indicate whether the 120 items are commonly used in English or not using a Yes/No button. The instructions for the task were adapted from Öksüz et al. (2020) and translated into Arabic by the researcher to ensure that the sampled intermediate L2 learners of English would understand the requirements of the task. Further, a slight change was introduced into the instructions. An additional sentence was added at the end of the instructions to inform the participants about the practice session. The practice session consisted of 12 items: three high-frequency collocations, three low-frequency collocations, and six baseline items. To minimize practice effects, the practice session items were adopted from Sonbul (2015) because the study looked at both high and low collocations. Although Sonbul (2015) did not create the task items based on their Log Dice scores and their MI scores, adopting Sonbul's items in the practice session was deemed acceptable as the goal of the practice is to familiarize the participants with the task.

3.3 Procedure

Both the proficiency measure and the main task were conducted online due to safety concerns during the era of COVID-19. First, the participants were given the consent form to understand the purpose of the research and indicate their willingness to participate in it. Then, the participants were asked to complete the web-based LexTALE test one week before administering the main task. This was done for two reasons: (a) to select a homogenous participant pool at the intermediate proficiency stage, and (b) to avoid any potential effects of the proficiency measure on the main task, such as participant fatigue. Only those who scored between 55 and 65 on the test were asked to do the acceptability judgment task (This decision was discussed in detail in the participant section above). This is done to ensure that the participants fall in the intermediate proficiency band, unlike the advanced learners in Öksüz et al. (2020), who had a mean score of 85 on the test (N = 32, *M* = 84.89, *SD* = 5.3).

A week after completing the LexTALE test, a link to the acceptability judgment task on the Gorilla platform was sent to the participants. The instructions of the task were in Arabic (see the Instrument section). The task was composed of four parts. (1) Instructions of the task were presented first, (2) the practice session, (3) the main session, and (4) finally a language background questionnaire. Following Öksüz et al. (2020), in both the practice and main sessions, a fixation point was presented (2,000 milliseconds) followed by the word pair remained on the screen until the participant makes a yes/no lexical decision or after a 5,000-milliseconds timeout. The order of the items was randomized across the participants to avoid potential order effects (e.g., later word pairs are answered correctly while earlier ones are not). The Gorilla platform recorded accuracy and response time for each participant in all the sessions. It took the participants approximately seven minutes to complete the acceptability judgment task. As a reward for task completion, the participants received a 16 SAR online voucher for a cup of coffee. To ensure that the data is gathered ethically (Mackey & Gass, 2016), an application for IRB approval was submitted, and informed consent from the participants was obtained.

4. Analysis

The analysis follows to a great extent Öksüz et al. (2020) study. To answer the research question, the response time of correctly accepted high-frequency collocations would be compared to their matched controls that were correctly rejected following previous research practice (Öksüz et al., 2020; Yi, 2018). Linear Mixed-Effect models (LME) using the lme4 package (Bates et al., 2015) were created on the R statistical platform (R Core Team, 2012). LME models have several advantages. They (1) analyze repeated data points (e.g., several reaction times from each participant), (2) allow including participants and items as random variables, which helps in controlling for variation in time reaction within participants and items (e.g., differences in reaction times between fast participants and slower ones and differences between easy and difficult items are accounted for) and (3) are robust to missing data points and unbalanced designs (Baayen, 2008). For these reasons, the present study used LME to analyze the participants' responses on the timed acceptability judgement task.

4.1 Data Trimming

Before doing the analysis, the reaction time data of the acceptability judgment task were trimmed. First, following previous research practice (Öksüz et al., 2020), incorrect responses were removed, such as when a collocation received a "no" response and when a control item received a "yes" response. The L2 participants correctly judged 83% of the high-frequency collocations and 75% of the low-frequency collocations. Meanwhile, only 58.3% of the baseline items were correctly judged (i.e., not accepted). Overall, correct responses constituted 68.7% (N = 1814) of total responses (N = 2640). A summary of correct and incorrect responses in the three conditions is presented in Figure 1.

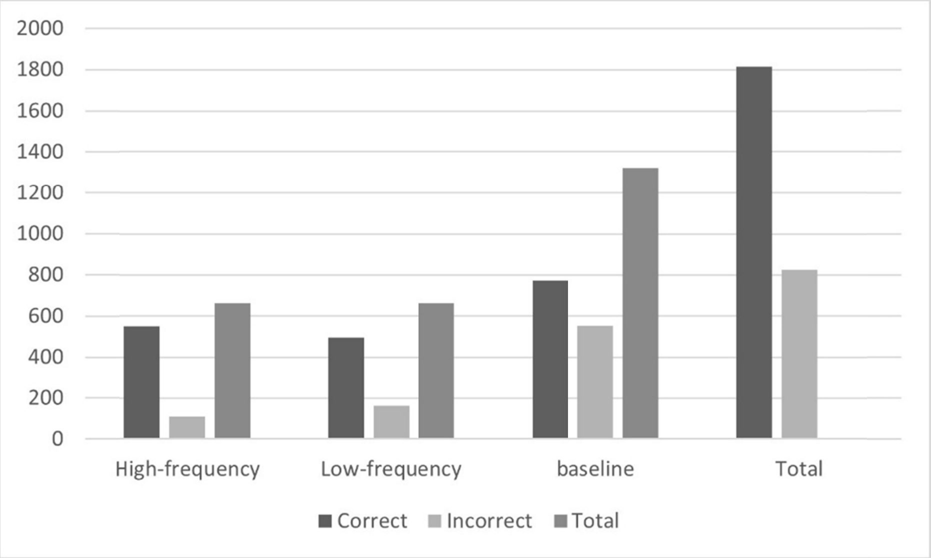

Figure 1. Summary of responses across the three conditions

Second, responses that were shorter than 450 milliseconds and responses that took longer than 5,000 milliseconds were excluded from the analysis following Öksüz et al. (2020) and Yi (2018). Thirty-nine responses (2%) were faster than 450 milliseconds; 38 of them were from the same participant. Fifty-nine responses (3%) were longer than 5,000 milliseconds. The removal of these short and long response times (RT) resulted in the loss of 5.4% of correct responses. The remaining responses after doing steps one and two amounted to 1715. Table 4 summarizes the mean response time of accurate answers for high-frequency, low-frequency, and baseline items in the acceptability judgment task.

Table 4. Mean response time in milliseconds across the three conditions

Condition	Mean	SD	Min	Max
High-frequency	1644.6	751	611	4886
Low-frequency	1728	675.8	642	4686
Baseline	2053	840.9	681	4931

4.2 Preliminary Analyses

Numerical and graphical analyses revealed that RT was non-normally distributed, with skewness of 1.2 (SE = 0.059) and kurtosis of 1.6 (SE = 0.118). The RTs were log transformed in R using the built-in log() function to reduce skewness in the distribution following prior research practice (Öksüz et al., 2020; Yi, 2018). The log transformed RT had a mean of 7.43 (SD = 0.40, Min = 6.41, Max = 8.50, bootstrapped 95% CI [7.41, 7.45]) and was normally distributed, with skewness of 0.24 (SE = 0.059) and kurtosis of -0.34 (SE = 0.118).

Also, the continuous predictors were mean-centered and standardized to allow comparison. This step was done because the examined continuous predictors are measured in different units, making it difficult to compare effects across the predictors. For instance, the LexTALE scores are out of 100, while the Log Dice scores are out of 16. A summary of the continuous variables is presented in Table 5.

Table 5. Summary of continuous predictor variables after centering and standardizing

Variable	Mean	Range	SD	Median
Collocation frequency	2.47	0.0, 4.93	1.30	2.65
Log Dice scores	2.91	-3.2, 10.95	3.74	3.0
MI scores	0.80	-5.15, 9.09	3.65	0.61
LeXTALE scores	0.00	-.93, 1.92	1.0	-0.49

Further, multicollinearity was checked before building the mixed-effect models. Multicollinearity is a potential issue that occurs when several independent variables are included in a mixed-effect model. Specifically, it describes the strong correlation between two or more of the independent variables. This generates inaccurate coefficient estimates and large standard errors, making it difficult to spot differences even when they exist. One of the most common ways to diagnose multicollinearity is by calculating the Variance Inflation Factor (VIF). VIF can be calculated using the function vif() in the car package in R (Fox & Weisberg, 2019). VIF values ≥ five are regarded by some practitioners as indicative of problematic multicollinearity, while others set the cutoff threshold to ≥ 10 (Levshina, 2015). The cutoff VIF score in this study was set at five following Öksüz et al. (2020).

Before creating the model, the VIF scores for each predictor variable entered in the model (collocation frequency, Log Dice scores, and MI scores) were calculated to avoid any potential multicollinearity problem. The VIF scores of collocation frequency, Log Dice score, and MI score were high (VIF = 35.2, 225.1, 107.2, respectively), suggesting a multicollinearity problem. When the Log Dice score variable was removed, the VIF scores of collocation frequency and MI score decreased to 5.3 and 5.2, respectively. This suggests that the two variables, Log Dice score and MI score, cannot be included in the same mixed-effects model because they would reduce the statistical power of the model when combined. In other words, two different models should be created to separately assess the influence of Log Dice score and MI score on collocation processing. Thus, this study built two mixed-effect models and compared the two models' predictive power using the Akaike Information Criterion (AIC) values of the models. This comparison is made to determine whether the model with the Log Dice scores or the one with MI scores better predict the participants' performance.

5. Results

The research question aimed to establish whether the Log Dice score is a better predictor of intermediate-level L2 English speakers' processing of collocations compared to Log Dice scores. To answer this question, two best fit models were built using the lmer4 package on R version 4.0.3. The analysis started with building simple models containing the fewest and most basic fixed and random effects (Larson-Hall, 2015). Then, fixed and random effects were added incrementally to the model to build the best fit model for the data. Introducing a small change to the model one at a time allows examining whether the added change is needed and enhances the model fit (Baayen, 2008). Thus, the old and modified models were compared in terms of their AIC scores using the built-in anova() command in R with each addition to the model.

The anova() function compares the AIC scores of the models and considers the model with the lower scores as the better fit because the model explains more variance in the data (Larson-Hall, 2015). To illustrate, this study first built a simple mixed-effects model with only the dependent factor (reaction time) and two random intercepts (items and subjects), which allows reaction time to vary across items and participants. Then, a second model was built, which had the same effects as the first model but included Log Dice scores and a fixed effect (an independent factor that explains variation in the dependent factor, i.e., reaction time). Using the anova () function, a comparison of the two models' AIC scores showed that the second model was the better fit (AIC = 1235) compared to the first model (AIC = 1273). This indicates that the addition of Log Dice scores to the model significantly improved model fit ($\chi^2(1) = 39.94$, p-value =0.00). As such, Log Dice scores were kept in the model as a fixed effect. This process was repeated with the inclusion of each new fixed and random effect until the best fit model was found, as indicated by its low AIC scores.

Using the model comparison approach, two best-fit models were built. One model examined the effects of LD scores on reaction time, while the other examined the effects of MI scores. Only correct responses were included in the two models following Öksüz et al. (2020). The two models had similar fixed/random effects. Both had participants/items as random variables and log-transformed reaction time as the dependent variable. The fixed effects in the two models were item type (high frequency, low frequency, baseline) and two item-related variables: noun frequency and item length. Also, Log Dice scores were entered as a fixed effect in the Log Dice model only, while MI scores were included only in the MI model as the two scores are collinear. The two models have only random intercepts as they explained some variance in the data, and no random slopes were included because they were redundant.

For the fixed effects, parameter estimates, 95% confidence intervals (CI) for those estimates, and p-values were reported. The CIs were calculated using the lmerTest package in R (Kuznetsova et al., 2017). For the random effects, variances and standard deviations are reported. Also, effect sizes for the two models are reported using marginal and conditional R^2 values, which highlight how much variation the model explains. The marginal R^2 considers only the fixed effects, whereas the conditional R^2 includes random effects as well. These values were calculated using the MuMIn package in R (Barton, 2019).

5.1 Log Dice Scores Effects on Reaction Time

The best-fit model for variables predicting reaction time in the Log Dice model is represented in Table 6. There are three significant fixed effects. First, item type had a significant effect on reaction time, the more frequent the collocation, the shorter the reaction time. Second, noun frequency influenced reaction time in that the more frequent the noun, the shorter the reaction time was. Third, item length had a significant effect on the reaction time measure; the longer the item, the longer the reaction time was. Finally, the Log Dice score did not have a significant effect on reaction time. This model explains 36% of the variance in reaction time as indicated by the R^2 conditional value, and this effect size is considered to be medium according to Plonsky and Ghanbar's (2018) guidelines. The R^2 conditional value is close to the one reported by Öksüz et al. (2020) [R^2 conditional = 0.31].

Table 6. Summary of the best-fit LME model for variables predicting L2 learners' RT data (N = 1715, R2 marginal = 0.087, R2 conditional = 0.36)

					Random effects			
	Fixed effects				By Subject		By Item	
Parameters	Estimate	SE	95% CI	T value (p)	Variance	SD	Variance	SD
Intercept	7.83	0.26	[7.32, 8.35]	29.8(***)	0.03	0.19	0.00	0.09
Log Dice score	-0.00	0.00	[-0.02, 0.00]	-1.10				
Item type	-0.15	0.04	[0.25, -0.06]	-3.17(**)				
Noun frequency	-0.09	0.04	[-0.18, -0.00]	-2.00(*)				
Item length	0.01	0.00	[0.01, 0.02]	4.25(***)				

Model formula: LogRT ~ LD_score + ItemType + noun_frequency + item_length + (1|Item) + (1|ParticipantID). *Item type was dummy coded, making baseline items as the reference group*. *p < .05; ** p < .01; *** p < .001.

5.2 MI Scores Effects on Reaction Time

Table 7 presents the best-fit LME model for the variables predicting reaction time in the MI model. Only two main effects were found to be significant: item type (the more frequent the item, the shorter the reaction time) and item length (the longer the item, the longer the reaction time). However, both MI score and noun frequency did not surface as significant main predictors as their 95% CIs went through zero. Like the Log Dice model, the R^2 conditional value for the MI model indicates that the model explains 36% of the variance in reaction time. Plonsky and Ghanbar's (2018) guidelines suggest that this effect size is medium.

Table 7. Summary of the best-fit LME model for variables predicting L2 learners' RT data (N = 1715, R2 marginal = 0.086, R2 conditional = 0.36)

					Random effects			
	Fixed effects				By Subject		By Item	
Parameters	Estimate	SE	95% CI	T value (p)	Variance	SD	Variance	SD
Intercept	7.82	0.26	[7.30, 8.34]	29.6 (***)	0.03	0.19	0.00	0.09
MI score	-0.00	0.00	[-0.01, 0.00]	-0.64				
Item type	-0.18	0.04	[-0.26, -0.00]	-4.17 (***)				
Noun frequency	-0.09	0.04	[-0.18, 0.00]	-1.96				
Item length	0.02	0.00	[0.01, 0.00]	4.31 (***)				

Model formula: LogRT ~ MI_score + ItemType + noun_frequency + item_length + (1|Item) + (1|ParticipantID). *Item type was dummy coded, making baseline items as the reference group*. *p < .05; ** p < .01; *** p < .001.

A comparison between the Log Dice and MI models was conducted using the anova() function in R. According to the AIC values, the model including the Log Dice scores was a better-fitting model a better-fitting model (AIC = 1211.9) than the model including MI scores (AIC = 1212.7). However, the value of the chi-square test for significance suggested no significant difference between the two models ($\chi^2(1) = 0$). This implies that the two models provide similar predictions.

6. Discussion

This study investigated the influence of an under-explored association measure (Log Dice scores) on L2 collocation reaction time. Specifically, this study examined the effects of two association measures (Log Dice scores and MI scores) on the processing of L2 adjective-noun collocations varying in frequency (e.g., long time, round face). The examined measures differ in how they identify collocations, with the MI-based method favoring low-frequency collocations. To examine any difference between the two measures, a timed acceptability judgment task was completed by 22 intermediate Saudi L2 learners of English. In this online-conducted task, the participants were asked to indicate whether (a) high-frequency collocations, (b) low-frequency collocations, and baseline items are commonly used in English using a Yes/No format. The data were analyzed using fixed-effects models on R as the observations were dependent (i.e., repeated measures from the same participant). Two best-fit models were created using the model comparison and testing method. One model was built for predicting the effects of Log Dice scores on reaction time and the other for MI scores. The decision to create two separate models was based on the observation that the two association measures were collinear (FIV = 225.1, 107.2, respectively) and thus cannot be included in the same model as they would mask each other's effect even if they exist.

The results of the two mixed-effect models show that the participants' reaction time was not affected by the type of association measure. Other factors determined the speed of processing L2 collocations for the sampled intermediate participants. Looking at the Log Dice score model results, there were three main predictors: item type, noun frequency, and item length. The participants' reaction time was shorter when the shown item had a higher frequency, had a more frequent noun, and had shorter component words (fewer characters). Likewise, for the MI scores model, other factors such as item type and item length were significant predictors of reaction time, while MI scores and noun frequency did not significantly affect reaction time.

Noun frequency was only a significant predictor in the Log Dice score model. One explanation for this might be the fact that the two association measures treat collocations differently. The Log Dice measure treats collocations in the same way, regardless of the frequency of their component words, while the MI prefers collocations with lower-frequency component words. For instance, one of the high-frequency collocations in the materials "long time" had a Log Dice score of 7.28, while its MI score was only 3.18. This collocation has a low MI score because the individual component words "long" and "time" frequently occur in the BNC corpus. In contrast, the Log Dice scores reward similar scores for collocations with higher- and lower-frequency component words and do not have a low-frequency bias like the MI measure. For this reason, noun frequency was a significant predictor in the Log Dice scores model but not the MI scores model.

Overall, the results of the two models suggest that the frequency of the collocation (operationalized as item type) and its length were the main factors influencing the intermediate participants' reaction time. This study determined the item type based on two aspects: (1) the raw frequency of the word pair in the BNC corpus, and (2) the pair's Log Dice score. Items with a frequency ≥ 10 and Log Dice ≥ 2 were considered as collocations, and word pairs with a lower frequency and Log Dice score were baseline items. The combination of raw frequency and Log Dice scores were a significant predictor as indicated by the significance of the item type factor in both the Log Dice and MI models. Interestingly, although item type, which is a composite of raw frequency count and Log Dice scores, was significant, Log Dice scores did not surface separately as a significant predictor. The fact that Log Dice scores and MI scores did not individually predict reaction time in the present study suggests that the participating intermediate learners did not yet develop sensitivity to collocation strength. The strength of collocation describes whether the word pair almost always appear together (Brezina, 2018). The participants seem to be at a proficiency range that does not allow them to capture collocation strength patterns in their L2 input.

The result that item type predicted the processing of collocations might imply that intermediate learners are sensitive to the frequency of the collocation. The participants' sensitivity to collocation frequency as indicated by the shorter time reaction to higher frequency collocations supports the usage-based model (Bybee, 1998, 2006; Tomasello, 2003), which proposes that language processing is influenced greatly by frequency of exposure. Recurrent language items in the input are processed and responded to faster, unlike those occurring less frequently. The usage-based model partly explains why the intermediate participants in this study recognized frequent collocations faster than less frequent baselines items. The current result may also indicate that learners become sensitive to frequency patterns in their L2 input as early as the intermediate proficiency stage.

The finding that MI scores did not predict collocation reaction time aligns with González Fernández and Schmitt (2015), who used a pen-and-paper collocation recall test. It should be noted that González Fernández and Schmitt (2015) and the present study differed in several ways. For example, while González Fernández and Schmitt tested the productive knowledge of collocations among L2 Spanish learners from three proficiency levels using an untimed task, the present study examined the receptive knowledge of collocations among L2 Arab learners in the intermediate level using a timed task. However, the result that MI scores did not influence reaction time contradicted Yi's (2018) findings, who used a timed acceptability judgment task and drawn materials from the BNC corpus. Nevertheless, Yi (2018) examined advanced L2 learners of English from L1 Chinese background. The difference in findings might be attributed to the fact that Yi (2018) sampled highly proficient L2 speakers, whereas the present study focused on intermediate L2 learners. Also, all the 32 L2 participants in Yi (2018) had lived in the United States for one year or more, suggesting extensive language exposure to native speech. Most of the L2 participants in the current study (91%) did not live in an English-speaking country.

On the other hand, the lack of significant effects for Log Dice and MI scores did not go in line with the findings of Öksüz et al. (2020). In Öksüz et al. (2020), advanced Turkish L2 learners of English were sensitive to both Log Dice and MI scores as indicated by the analysis of their reaction time in a timed acceptability judgment task. Although the current study adopted the materials used in Öksüz et al. (2020) and replicated the task on a different L2 group (intermediate Arab learners), the present study' findings did not support what was reported in Öksüz et al. (2020).

Three reasons could explain this divergence in findings between the present study and Öksüz et al. (2020). First, the L2 group recruited in Öksüz et al. (2020) differed from the participating group in the present study in terms of proficiency level and level of exposure to the L2. The mean LexTALe scores for the L2 learners in Öksüz et al. (2020) was 85 (BCa 95% CI = [83, 86.7]), while it was 52.6 in the current study (BCa 95% CI [54.5, 58,7]). Furthermore, out of the 32 L2 participants in Öksüz et al. (2020), 21 (65.6%) had stayed in an English-speaking country for more than one month, while only 2 (9%) of the 22 participants from the present study did so. Second, the materials were drawn from the BNC XML corpus, and it is possible that the language experience of the L2 participants in the present study was different from the one represented in the corpus. For instance, several high- and low-frequency items drawn from this corpus might not be familiar to those who did not live in an English-speaking country or are not well-read in politics, such as "*common law, labor party, social policy, local government, special court*". While the BNC XML corpus might not have reflected the language experience of the L2 participants in the current study as each one of them was exposed to a specific type of English, it might have matched the advanced L2 group's experience with English recruited in Öksüz et al. (2020) since more than half of them lived in an English-speaking country.

The third point that could explain the difference between the two studies is that they included different fixed effects. The best-fit model reported in Öksüz et al. (2020) included three predictor variables ((a) L1 vs. L2, (b) MI/Log

Dice scores, (c) group*MI/Log Dice scores) while the present study included four ((a) MI/Log Dice scores, (b) item type, (c) noun frequency, (d) item length). As can be seen, the present study and Öksüz et al. (2020) included different fixed effects except for MI/Log Dice scores. The analysis in the present study aimed to build the best-fit model with the lowest AIC scores and used the model comparison and testing method to do so. As the present study used this method to build mixed-effect models, this resulted in including as many significant fixed effects as possible to account for the participants' reaction time. It was possible to follow the model formula used in Öksüz et al. (2020), but the outcome model would have higher AIC which would explain smaller percentage of the variance in reaction time (i.e., weaker model with limited explanatory power). Thus, the present study used the model comparison method to build the best-fit model which resulted in including more and different fixed effects compared to Öksüz et al. (2020).

Overall, the present study results suggest that intermediate English L2 learners are not sensitive to corpus-based association measures. One important characteristic of the sampled participants is that most of them (91%) studied the L2 in their home country and had not lived in an English-speaking country. This suggests that L2 learners might not be attuned to corpus-based association measures at intermediate levels, specifically if they did not live abroad. Being sensitive to such measures possibly requires reaching higher proficiency levels.

These results have important implications for L2 teaching and testing. It might not be worthwhile to determine which collocations to include in the materials based mainly on the strength of the association. Learners at the intermediate stage do not yet seem to pay much attention to collocation strength when processing (e.g., reading) collocations. At this proficiency level, learners have not yet been fully developed the skill to notice whether the component words of a collocation appear almost always together. For instance, they seem not to be greatly sensitive to the fact that the adjective "only" and the noun "friend" co-occur together in the same order in many contexts. Therefore, it is not preferred to use one collocation association measure as the sole criteria for constructing teaching materials. Rather, building collocation teaching materials should be informed by several characteristics of the collocation. One characteristic is the semantic transparency of the collocation. For instance, the collocation "tighten strings" has a transparent meaning in that its meaning can be understood by looking only at the literal meaning of the component words, while this is not true for "pull strings". Another collocation characteristic that should be considered when choosing appropriate materials is the frequency of the occurrence. Since item type was a significant predictor in the analysis, teaching intermediate learners collocations with both high raw frequency and Log Dice scores might be practical. Also, initially, collocations with shorter component words might be introduced as they are easier to process for those in the intermediate range. Gradually, longer collocations with a high frequency of occurrence can be presented to learners to familiarize them with this type of collocations. As for testing collocation knowledge, the present study results suggest that testing intermediate learners' knowledge of collocations based only on their high MI scores may not be the most appropriate method. It is more valid to include collocations with both high raw frequency counts and Log Dice scores in tests assessing L2 collocation recognition ability. Learners in the intermediate range are more likely to be familiar with this kind of collocations.

7. Limitations and Future Research

This study is limited in several ways. The recruited sample is small compared to previous studies, and the results might not be as generalizable. Only females participated in the study to ensure similarity of L2 instruction, and thus the results might not apply to Arab male learners. Also, only one proficiency level was examined. The analysis of different levels would have revealed a fuller picture of the effect of association measures on L2 collocation processing.

Further, the study did not conduct an accuracy analysis of the participants' responses on the grammaticality judgment task because the main aim was to investigate the participants' speed of processing (reaction time). Another limitation is that the study focused mainly on adjective-noun collocations and did not examine other types of collocations (e.g., verb-noun, preposition-noun, adjective-preposition), and the conclusions of the present study cannot extend to all collocation types. Additionally, the task used in the study gauges the receptive knowledge of collocations. As a result, the effect of corpus association measures on the productive knowledge of collocations was not examined. Further, L2 proficiency was measured using a vocabulary test rather than a standardized measure that assesses the four language skills. This way of estimating proficiency is less comprehensive and accurate because the participants' proficiency level might not be fully captured by their scores on the vocabulary test.

Future research could examine a larger sample of Arab L2 learners of English and include both females and males to evaluate the generalizability of the present study. In addition, the effect of association measures should be examined in terms of both accuracy of response and reaction time to better understand such an effect. Likewise, the use of an additional task to examine the productive knowledge of collocation along with the grammaticality judgment task would help in understanding the influence of association measures.

8. Conclusion

This study examined the effects of two corpus-based association measures on the processing of collocations among intermediate L2 learners of English. While works on L2 collocational processing relied largely on MI scores to extract collocations, limited research has assessed the suitability of such an association measure in gauging L2 learners' knowledge of collocations. Following Öksüz et al. (2020), the present study aimed to determine whether MI is a stronger predictor of L2 performance compared to the Log Dice measure. The timed acceptability judgment task revealed that learners at the intermediate stage do not seem to be sensitive yet to corpus-based association measures. Other factors influenced their performance on the task, such as the frequency of the collocation (operationalized as item type) and its length. The results have important implications for L2 teaching and testing and may indicate that it is not worthwhile to determine which collocations to include in the materials based mainly on the strength of the association.

Acknowledgments

The author would like to thank the anonymous reviewers and editors for their insightful and helpful comments. The author would also like to express her gratitude to Abdulmajeed Alzahrani for his supportive feedback.

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

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2019). Gorilla in our Midst: An online behavioral experiment builder. *Behavior Research Methods*, 1–20. https://doi.org/10.1101/438242

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press. https://doi.org/10.1017/CBO9780511801686

Barton, K. (2019). *MuMIn: multi-model inference*. R package version 1.43.17 [Computer software]. Retrieved from https://cran.r-project.org/web/packages/MuMIn/index.html

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting-linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, *69*, 65–78. https://doi.org/10.1016/j.system.2017.08.004

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28–41. https://doi.org/10.1016/j.jslw.2014.09.004

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press. https://doi.org/10.1017/9781316410899

Bybee, J. (1998). The emergent lexicon. *Chicago Linguistic Society*, *34*, 421–435.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 711–733. https://doi.org/10.1353/lan.2006.0186

Cowie, A. P. (1998). Introduction. In A. P. Cowie (Ed.), *Phraseology: Theory, Analysis, and Applications* (pp. 1–20). Oxford University Press.

Davies, M. (2004). *BYU-BNC: The British National Corpus. 100 million, British, 1980s–1993*. Retrieved from http://corpus.byu.edu/bnc

Davies, M. (2008). *The Corpus of Contemporary American English (COCA): 1.0 billion, US, 1990–2019*. Retrieved from https://www.english-corpora.org/coca/

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, *28*(3), 157–169. https://doi.org/10.1016/j.esp.2009.02.002

Ellis, N. C., Simpson‐Vlach, R. I. T. A., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *Tesol Quarterly*, *42*(3), 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Evert, S. (2008). Corpora and collocations. In A. Ludeling (Ed.), *Corpus linguistics: An international handbook* (pp. 1212–1248). Mouton de Gruyter.

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage publications.

Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P., & Sharma, N. (2019). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, *31*(1), 23–39. https://doi.org/10.1017/S0958344018000150

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus‐based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, *67*, 155–179. https://doi.org/10.1111/lang.12225

González Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have?: The effects of frequency and amount of exposure. *International Journal of Applied Linguistics*, *166*, 94–126. https://doi.org/10.1075/itl.166.1.03fer

Gries, S. T. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics*, *18*(1), 137–166. https://doi.org/10.1075/ijcl.18.1.09gri

Henriksen, B. (2013). Research on L2 learners' collocational competence and development: A progress report. In C. Bardel, C. Lindqvist & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 29–56). Eurosla.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS*. Routledge. https://doi.org/10.4324/9781315775661

Laufer, B., & Waldman, T. (2011). Verb‐noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*(2), 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*, 325–343. https://doi.org/10.3758/s13428-011-0146-0

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins. https://doi.org/10.1075/z.195

Mackey, A., & Gass, S. M. (2016). *Second language research: Methodology and design* (2nd ed.). Taylor & Francis. https://doi.org/10.4324/9781315750606

McCauley, S. M., & Christiansen, M. H. (2017). Computational investigations of multiword chunks in language learning. *Topics in Cognitive Science*, *9*, 637–652. https://doi.org/10.1111/tops.12258

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. John Benjamins. https://doi.org/10.1075/scl.14

Nguyen, T. M. H., & Webb, S. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research*, *21*(3), 298–320. https://doi.org/10.1177/1362168816639619

Öksüz, D., Brezina, V., & Rebuschat, P. (2020). Collocational Processing in L1 and L2: The Effects of Word Frequency, Collocational Frequency, and Association. *Language Learning*, *71*(1), 55–98. https://doi.org/10.1111/lang.12427

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal*, *102*(4), 713–731. https://doi.org/10.1111/modl.12509

R Core Team. (2012). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Rychly, P. (2008). A lexicographer-friendly association score. In P. Sojka & A. Hor'ak (Eds.), *Proceedings of* *recent advances in Slavonic natural language processing* (pp. 6–9). RASLAN.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Sonbul, S. (2015). Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off‐line/online collocational processing. *Bilingualism: Language and Cognition*, *18*, 419–437. https://doi.org/10.1017/S1366728914000674

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics*, *32*(1), 231–254. https://doi.org/10.1017/S026719051200013X

Xu, J. (2018). Measuring "Spoken Collocational Competence" in Communicative Speaking Assessment. *Language Assessment Quarterly*, *15*(3), 255–272. https://doi.org/10.1080/15434303.2018.1482900

Yi, W. (2018). Statistical Sensitivity, Cognitive Aptitudes, and Processing of Collocations. *Studies in Second Language Acquisition*, *40*(4), 831–856. https://doi.org/10.1017/S0272263118000141

Copyrights