Applied Psycholinguistics (2020), **41**, 1197–1222
doi:10.1017/S014271642000051X

CAMBRIDGE UNIVERSITY PRESS

ORIGINAL ARTICLE

Timed versus untimed recognition of L2 collocations: Does estimated proficiency modulate congruency effects?

Suhad Sonbul[1,*] and Dina El-Dakhs[2,*]

[1]Umm Al-Qura University and [2]Prince Sultan University
*Corresponding author. E-mails: sssonbul@uqu.edu.sa; ddakhs@psu.edu.sa.

(Received 4 August 2019; revised 3 July 2020; accepted 27 August 2020)

Abstract

Congruency (the availability of a direct first language translation) and level of proficiency have been reported among the most important determinants of second language collocation processing. However, only very few studies looked at the interaction between the two determinants, and none of these directly compared untimed collocation recognition assessed through traditional tests to timed recognition evident in psycholinguistic tasks. The current study administered both types of form recognition measures to 228 female Saudi English as a foreign language learners in two separate experiments: a traditional multiple-choice test (Experiment 1) and a timed acceptability judgment task (Experiment 2). Experiment 2 also tested 37 native speakers of English as a baseline for comparison. Congruency, estimated proficiency (vocabulary test scores), and the interaction between the two were evaluated as predictors of untimed and timed recognition through mixed-effects modeling. Results showed that congruency and estimated proficiency had a clear effect on untimed and timed recognition. More interesting, the effect of proficiency was clearer on timed recognition with a gradual decrease in the first language effect as proficiency increased getting closer to nativelike collocation processing. Results have implications for second language collocation learning and testing.

Keywords: collocations; congruency; estimated proficiency; timed and untimed recognition; vocabulary

A central component of lexical knowledge is formulaic language. Research has indisputably shown formulaicity as a widespread component of language use (Erman & Warren, 2000; Foster, 2001). Sounding native does not only entail knowledge of grammatical rules but also the ability to use words appropriately in acceptable (formulaic) sequences including collocations (Pawley & Syder, 1983). However, second language (L2) learners do not always reach that nativelike level. Studies evaluating English as a foreign language (EFL) learners' knowledge of collocations have generally highlighted the difficulties they face in recognizing and/or recalling acceptable collocation forms (e.g., El-Dakhs, 2015; Farghal & Obeidat, 1995; González Fernández & Schmitt, 2015; Hussein, 1990) with a clear

influence from their first language (L1; Bahns & Eldaw, 1993; Biskup, 1992; Nguyen & Webb, 2017). A similar L1 effect was also reported in several corpus-based studies that examined EFL learners' use of collocations (e.g., Laufer & Waldman, 2011; Nesselhauf, 2003).

More recently, L1–L2 collocation congruency, that is, the availability of a direct translation match between the two languages, has attracted increasing research attention. In contrast with early studies that employed traditional tests, this current line of research is led by Brent Wolter, Junko Yamashita, and Henrik Gyllstad (e.g., Wolter & Gyllstad, 2013; Yamashita & Jiang, 2010, see below) who employed timed psycholinguistic tasks that are assumed to reflect automatic language processing (Ellis, 2005).

Although interesting and informative, these studies are limited in several ways. Only one study (Wolter & Gyllstad, 2011) employed measures of both timed and untimed recognition to allow for a comparison between them. More importantly, the effect of proficiency level (lower vs. higher) on sensitivity to L1 congruency during timed recognition was explored directly only in a few studies mainly in the Japanese context (e.g., Wolter & Yamashita, 2018; Yamashita & Jiang, 2010) with contrasting results. Finally, the available research has mainly focused on Japanese and Swedish learners of English. Congruency effects might vary according to the learners' L1 and its proximity to the L2. The present study aims to address these limitations. In two experiments, it employs two of the most widely used recognition measures of untimed (multiple-choice test) and timed (acceptability judgment task) collocation recognition with Arab EFL learners in order to explore the contribution of L1 congruency, estimated L2 proficiency level, and the interaction between the two. The study further adds to earlier research through dealing with proficiency as a continuous rather than a categorical variable to allow for a clearer depiction of the interaction between L2 ability and congruency. The findings can help better understand the nature of L2 collocation development and can, thus, lead to tangible implications for L2 pedagogy and assessment.

Background

Ellis (2005) distinguishes between controlled and automatic language processing. The former is often assessed through untimed measures while the latter is evident in timed performance. Recent eye-tracking evidence (Godfroid et al., 2015) has shown that the two forms of processing are distinct. In the present study, we will adopt Ellis's (2005) terms in referring to measures of collocation processing.[1] Both types of processing can be assessed using either recall or recognition measures. Recall measures include translation, gap-fill, and naming tasks where participants need to generate a response from memory based on a presented cue. Recognition measures, in contrast, comprise multiple choice (MC), matching tests, or acceptability judgment/lexical decision tasks where the learner demonstrates the ability to recognize a presented item. The focus of the present study is on recognition of collocations (i.e., two-word pairs that occur more often than would be expected by chance). Thus, we define collocations following the frequency-based approach (see Gablasova, Brezina, & McEnery, 2017, for approaches to defining collocations).

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

In two early studies, Biskup (1992) and Bahns and Eldaw (1993) explored the effect of congruency on the ability of L2 learners to render acceptable collocations in traditional recall tasks. Biskup (1992) compared the errors made by 34 Polish and 28 German advanced learners of English in translating 23 collocations in an L1–L2 task and found that language distance made a difference with more L1-based errors for the former group. Similarly, Bahns and Eldaw (1993) found that their 58 German university-level learners of English made around 50% errors in an L1–L2 translation task and a cloze task involving 15 collocations, with a clear influence from the L1. In a more recent and comprehensive study, Nguyen and Webb (2017) employed a recognition MC task to examine the influence of node word frequency, collocation frequency, mutual information score (a measure of collocation strength), congruency, and word class on 100 Vietnamese EFL learners' recognition knowledge of 180 verb–noun (VN) and adjective–noun (AN) collocations at the first three 1,000-word frequency levels. Knowledge of single-word items at the same word frequency levels was also examined. Results showed low levels of collocation knowledge (around 45% correct answers) overall with node word frequency being, surprisingly, a stronger predictor of collocation knowledge than collocation frequency. Other less influential predictors of collocation recognition were mutual information score and congruency. However, word class (VN vs. AN) was not a significant predictor. Finally, a positive correlation was reported between collocation and single-word knowledge explaining around 45%–50% of the variance. This seems to suggest that as vocabulary size increases, which can be considered a rough estimate of proficiency (see Alderson, 2005; Gyllstad, 2007, 2009), so does knowledge of collocations. The study did not, however, attempt to test for the interaction between congruency and proficiency. Thus, although the three studies above are interesting in that they establish the effect of congruency on untimed collocation recognition, they did not explore the interaction between this L1 influence and L2 proficiency level. In addition, they did not evaluate timed recognition, which is assumed to reflect actual language use.

Wolter and Gyllstad (2011) employed a timed primed lexical decision task (LDT) in addition to an untimed test (Gyllstad's [2007] Yes/No COLLMATCH). A group of 35 EFL Swedish learners and 30 native speakers of English completed the LDT with verbs as primes and nouns as targets, and their reaction time (RT) to the targets was measured. Thus, the timed measure is assumed to tap into online processing as it unfolds in real time. The items were balanced for three types of VN pairs: congruent/incongruent collocations and noncollocate pairs, with 33 items under each category. Then, only the EFL learners took the Yes/No test. The results showed that the L1 may have considerable influence on both untimed and timed collocation processing. EFL learners responded to congruent collocations faster in the LDT and with higher recognition scores in the untimed test than incongruent collocations. As expected, natives, in contrast, did not show such an advantage in processing. However, some variability was noted for the incongruent collocations, which led to an insignificant by-item analysis in the LDT. It should be noted, however, that Wolter and Gyllstad (2011) only tested advanced nonnatives and, thus, did not clearly explore the effect of proficiency level on congruency effects.

This effect was the focus of Yamashita and Jiang's (2010) study. They investigated the effect of L1 congruency on the timed processing of VN and AN

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

collocations (24 congruent and 24 incongruent) by Japanese learners of English in different L2 contexts, English as a second language (ESL; $N = 24$) and EFL ($N = 23$), with a clear difference in proficiency according to scores of a cloze task. Their timed task was different from that employed by Wolter and Gyllstad (2011) in that it required an explicit judgment on the pair presented. This study also included a group of native English speakers ($N = 20$). Results showed that (a) both EFL and ESL learners made more errors on incongruent than congruent collocations, (b) only lower level EFL learners showed an RT advantage for congruent collocations, and (c) natives did not show such effects. It was concluded that even with a substantial amount of L2 exposure in the ESL context, incongruent collocations might be more difficult to learn, but once they are learned, they seem to be processed similarly to congruent collocations. It is noteworthy, though, that this study is limited in that the analysis did not include control pairs, which play an important role in setting a baseline for comparison.

In a similar study, Wolter and Yamashita (2018) employed a timed acceptability judgment task with 27 natives and 47 Japanese EFL learners but included four types of AN pairs: 24 congruent collocations, 24 incongruent collocations, 24 Japanese-only collocations (not possible in English; see also Wolter & Yamashita, 2015), and 72 control noncollocate pairs. In addition to evaluating congruency and L1 transfer effects, this study also investigated L2 proficiency (intermediate and advanced learners, $N = 24/23$, respectively) and word/collocation frequency. Deviating from previous research using the acceptability judgment task, this study analyzed RT of all "Yes" responses regardless of accuracy and did not analyze error rates. As for the frequency effect, all participants showed sensitivity to both word-level and collocation-level frequency with clearer effects for natives and the advanced EFL group than the intermediate-level group. In addition, a clear congruency effect was reported among EFL groups with faster processing for congruent collocations. Finally, RTs for Japanese-only pairs were not significantly different than control pairs.

In reconciling differences between their findings and those reported by Yamashita and Jiang (2010) regarding proficiency effects, Wolter and Yamashita (2018) pointed out that certain component words were repeated more than once in Yamashita and Jiang's (ibid.) study, creating a certain level of noise in the data. We believe that another important source of this variation in results might be the range of proficiency explored. Yamashita and Jiang's (2010) participants represented a sharp distinction in exposure (EFL vs. ESL context). In contrast, both groups of nonnatives in Wolter and Yamashita (2018) were close-level EFL learners impeding the establishment of any effect (see Carrol, Conklin, & Gyllstad, 2016, for a similar finding for L2 idioms). Only a few collocation studies included proficiency as a continuum rather than a categorical measure.

This was done by Wolter and Gyllstad (2013) who mainly focused on frequency effects on collocation processing. The items included 40 congruent and 40 incongruent AN collocations varying in corpus-derived frequency in addition to 80 control, noncollocate, pairs. Two groups of participants completed a timed acceptability judgment task: 25 Swedish advanced EFL learners and 25 native speakers of English. The results demonstrated a clear frequency effect on L2 collocation processing for both groups. The study also showed that EFL learners were sensitive to L1 congruency both in terms of accuracy (error rates) and fluency (RT). In addition, a post hoc

analysis with the nonnative data only showed that proficiency (as indicated by raw vocabulary size scores included as a continuous variable) was a major predictor of both RT and accuracy scores. More interestingly, the interaction between proficiency and congruency was only present for accuracy but not RT data. This result seems to suggest (similar to Wolter & Yamashita, 2018) that proficiency does not modulate speed of collocation access. However, again, it should be noted that variation in proficiency is fairly limited in this study with all participants being advanced leaners. This may have concealed differences in sensitivity to congruency between lower level and higher level nonnatives. Another important limitation of this study (and previous similar research) is that the control pairs devised for the timed measures were mostly semantically implausible (e.g., **sorry body*, **married top*; Wolter & Gyllstad, 2013) and might, thus, have put them at an inherent disadvantage in comparison to target collocations (see Sonbul & Siyanova-Chanturia, in press, for a lengthy discussion).

To wrap up, findings related to congruency effects on L2 collocation processing on both the untimed and timed sides of processing are far from being conclusive. While studies employing traditional measures (translation, cloze, and MC) have shown a clear effect of congruency on test scores, none of them explored the interaction between congruency and proficiency. As for timed processing, most of the studies reviewed manifested an L1 effect. However, only three studies (Wolter & Gyllstad, 2013; Wolter & Yamashita, 2018; Yamashita & Jiang, 2010) attempted to explore the interaction between congruency and proficiency with contradictory findings due to variation in gaps between participant groups. Moreover, only two English-language learning contexts (Japanese and Swedish) were explored limiting generalization to other L2 settings. Finally, control items in the available research were mainly semantically implausible.

Present study: Aim and research questions

The present study aims at dealing with the above-mentioned limitations through administering one traditional/untimed test (MC, Experiment 1) and another timed (acceptability judgment, Experiment 2) task to evaluate the form recognition ability of Arab EFL learners. Level of proficiency was estimated through scores in two levels of the Vocabulary Levels Test (VLT; Webb, Sasao, & Ballance, 2017) and was dealt with as a continuous measure. The study involved both congruent and incongruent collocations (both experiments) in addition to control pairs (Experiment 2 only). Item selection followed a different approach than previous research whereby control items are semantically plausible. The following research questions will be addressed:

1. Does congruency have an effect on untimed (Experiment 1) and timed (Experiment 2) collocation form recognition?
2. Does estimated L2 proficiency have an effect on untimed (Experiment 1) and timed (Experiment 2) collocation form recognition?
3. Does estimated L2 proficiency modulate the effect of congruency on untimed (Experiment 1) and timed (Experiment 2) collocation form recognition?

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

Based on findings from previous research, we predict congruency and estimated proficiency to have an effect on both timed and untimed collocation recognition. Thus, a positive answer is expected to both Research Questions 1 and 2. However, it is difficult to predict the interaction between congruency and proficiency (Research Question 3). Previous research seems to suggest that, on the timed side of processing, congruency effects are not modulated by proficiency when defined at a narrow scale or in categorical terms. It would be interesting to explore this interaction when proficiency is dealt with as a continuous variable in both untimed (Experiment 1) and timed (Experiment 2) tasks.

In addition to the main effects of congruency and proficiency, the present study also attempts to explore other factors which have been previously reported to influence collocation processing. These include word class (AN vs. VN collocations), pair frequency, and frequency of constituent words. Based on previous research employing timed tasks, we expect word and collocation frequency to show a clearer effect in Experiment 2. Word class is especially interesting in the present context due to variation in configuration between VN and AN collocations in English and Arabic: similar for VN collocations, with verbs preceding objects, but different for AN collocations, with nouns preceding adjectives in Arabic. Due to the mismatch in word order for AN collocations, one might expect them to exhibit slower processing in comparison to VN pairs by Arabic EFL learners.

Experiment 1

In this experiment, Arab learners of English are administered an MC test of collocations. The aim is to examine the effect of congruency and L2 proficiency on untimed collocation recognition and whether there is any interaction between the two factors.

Methods

Participants

One hundred and twenty-two female Arab learners ranging in age between 18 and 25 ($M = 20.91$, $SD = 1.82$) took part in this experiment. They were all BA students at a university in Saudi Arabia where English is the main medium of instruction and were recruited from various levels of study in order to ensure variability in English proficiency levels. Nearly half of them ($N = 60$) were students at the preparatory-year English language program while the other half ($N = 62$) were senior students who were studying in various majors after meeting their English-language requirement. All participants started learning English in schools at an average age of 8.30 years ($SD = 4.41$), and none had spent more than 3 months in an English-speaking country.

The two groups of participants were given the 1,000 (1k) and 2,000 (2k) levels of the Updated VLT, Version A (Webb, Sasao, & Ballance, 2017) as a rough estimate of proficiency. We measured the 1k and 2k levels only due to time limitations. Their average score out of 30 was 27.74 at the 1k level ($SD = 2.57$) and 22.77 at the 2k level ($SD = 5.85$). The total mean score across both levels was 50.51 ($SD = 7.92$). The Kuder–Richardson (Formula 21) reliability coefficient (see Bachman, 2004) for the VLT was found acceptable (0.89).[2] The distribution of total VLT scores is presented in Figure 1. It should be noted that these were centred prior to the analysis.

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

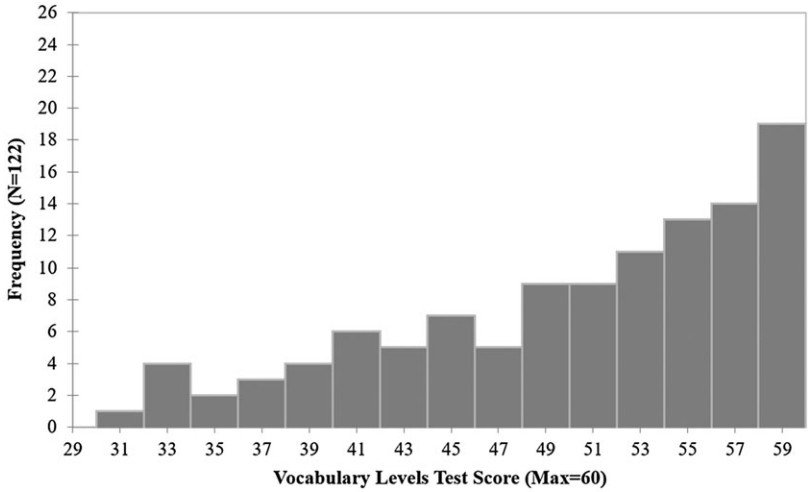

Figure 1. Distribution of VLT scores achieved by participants (Experiment 1).

Following Sonbul (2015), we included VLT scores as a continuous predictor variable in the mixed-logit model (see the Analysis section below). This step was taken to ensure a fine-tuned depiction of proficiency effects on collocation recognition.

Items

We started with a large cohort of AN and VN collocations from which the final set of target items were selected. In the initial stage of item pooling, we consulted various resources (Alharbi 2017; Migdad 2012; *Oxford Collocations Dictionary*, 2002). All candidate collocations were checked to make sure that the following criteria are met:

a. Component words (adjectives, nouns, and verbs) belong to the most frequent 2,000 word families in English according to Nation's (2012) BNC/COCA list. This step was important to ensure that our participants are likely to be familiar with the individual words comprising the target collocations (see Zhang, 2017).
b. Following Nguyen and Webb (2017), all collocations had a frequency of at least 50 in the Corpus of Contemporary American English, COCA (Davies, 2008) with a minimum mutual information (MI) score of 3. The window for AN collocations was set to –1 but the window for VN collocations was set to –2 to allow for an intervening determiner.

This initial step produced a long list of 968 AN collocations and 392 VN collocations. These items were then individually checked by the authors (both proficient L1 Arabic–L2 English bilinguals) to classify them either as congruent (L1 = L2, e.g., *natural birth*, where translating the core meaning of the individual words renders an acceptable collocation in Arabic) or as incongruent (L1 ≠ L2, e.g., *fresh start*, where the literal translation of core meanings is not acceptable in Arabic; an

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

equivalent of *new start* is used instead). This resulted in a list of collocations whose classification was shared by both authors: 34 congruent AN collocations, 28 incongruent AN collocations; 33 VN congruent collocations and 30 incongruent VN collocations. In order to validate this initial classification decision, all 125 candidate collocations were intermixed in a list and presented to four independent judges who hold MA or PhD degrees in Arabic–English translation. They were instructed to translate each English collocation into Arabic and then make a decision as to whether they (a) believe it is congruent (i.e., the literal translation of the English collocation is acceptable in Arabic), (b) believe it is incongruent (i.e., the literal translation of the English collocation is not acceptable in Arabic), or (c) are not sure. Only items with congruency/incongruency decision that is shared by at least three out of the four judges were classified as congruent or incongruent, respectively. This resulted in a final list of 40 target collocations with 10 target items under each category. None of the individual words comprising target collocations was repeated more than once in the list. This was important to avoid any undesirable repetition effect on test results. Appendix S.1 in the online supplemental material presents the full item list along with distractors for the MC test (see the Measure section below).

Item-related variables including individual word frequency, collocation frequency, and collocation length were not experimentally controlled for. As our analysis used mixed-logit models, we statistically controlled for these variables by including them as fixed factors in the model (see the Analysis section below).

Measures

We employed the most widely used collocation recognition measure, that is, MC. Three distractors were developed for each target item that met the following criteria:

1. All distractors belonged to the most frequent 2,000 word families in English (Nation, 2012).
2. Following Nguyen and Webb (2017), the distractor combined with the noun node (in a window of -1 for adjective and -2 for verbs) had a frequency between 0 and 10 in the COCA. When the pair occurred in the COCA, the MI score was either negative or 1 maximum signaling no association. The only exception to this is the distractor **push time* which has a frequency of 11 in the COCA but with a negative MI score (–3.20).

We then ensured that all distractors were semantically plausible. In a norming task, the initial version of the test was presented to 5 native speakers of English who hold MA or PhD degrees either in Linguistics or in TESOL. They were instructed to make any changes they deemed necessary in relation to the most suitable determiner in VN collocations. In addition to this, we asked them to mark any distractor that was particularly semantically marked (implausible). Any distractor that was marked as implausible by three native speakers was replaced by another one. Here are examples for the items *limited ability* and *catch cold*:

- _____ ability
  A. narrow B. gentle C. limited D. big

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

- _____ a cold
  A. ignore B. happen C. catch D. attract

Items were divided into two sections, AN and VN. The final set of 40 multiple-choice items was piloted with 5 native speakers of English (different from those who took the norming task above). They all achieved a perfect score (100%).

Scoring the test was straightforward. A response was scored 1 when it was correct and 0 if it was incorrect or missing. The reliability of the test was calculated based on the scores achieved by the 122 EFL participants and was found high (Cronbach $\alpha = 0.92$).

Procedures

The participants were recruited during their normal classes.[3] They took the MC test (around 20 min) followed by the 1k and 2k levels of the Updated VLT (around 15 min), and finally the language background questionnaire.

Analysis

The analysis was done using mixed-logit models for binary data with the *glmer* function (*lme4* package) on R version 3.5.0. (R Core Team, 2018). It was conducted in both the forward and backward methods to check consistency, and the final models were identical in both directions.[4]

The analysis included participants and items as random variables, 0/1 MC scores as the dependent measure, and the following fixed factors: item type (congruent vs. incongruent), VLT scores as a continuous variable representing estimated proficiency, and the interaction between the two main factors. We also tested for the potential effect of item-related variables on test scores. These included word class (AN vs. VN), word 1 frequency, word 2 frequency, pair frequency, and pair length. All frequencies were taken from the COCA and were log transformed to reduce skewness in data. Odds ratios (Exp[β]) transformed from log odds were used as effect size measures in the model (see Puimège & Peters, 2019).[5]

Collinearity between item-related variables was checked prior to the analysis, and residuals were computed for pair length and log word 2 frequency. Both residuals correlated well with the original variables (pair length $r = .92$, $p < .001$, and log word 2 frequency $r = .71$, $p < .001$). Table 1 summarizes continuous variables.

Results and discussion for Experiment 1

Table 2 below presents the number and percentage of correct/incorrect responses in the MC test. It can be clearly seen that, overall, our Saudi EFL learners scored higher for congruent than incongruent collocations (with a 11.23 and 12.13 percentile point difference for AN and VN collocations, respectively). In contrast, the advantage for AN over VN collocations is not large with a 1.31 percentile-point difference only for congruent and 2.21 percentile-point difference for incongruent collocations.

Table 3 presents the best-fit model for variables predicting MC scores. Only two main predictors were found significant. Estimated L2 proficiency as reflected by

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

Table 1. Summary of continuous variables (MC test)

Variable	Range (adjusted range)	*SD*	*Mdn*
Log pair frequency	4.21–8.18 (–1.60 – 2.37 log units)	1.07	0.06
Resid pair length	9–23 (–6.06 – 7.97 characters)	2.89	–0.79
Log word 1 frequency	8.75 –13.14 (–2.17 – 2.22 log units)	0.88	–0.14
Resid log word 2 frequency	9.19–13.72 (–2.11 – 1.30 log units)	0.63	0.11
VLT score	31.00–60.00 (–19.51 – 9.49 points)	7.88	2.49

Note: The second column shows the range of the variables. The adjusted range after transformation and/or centring is presented in parentheses. Standard deviations and medians refer to the predictor values in the models. All variables are centred, and their means are zero.

Table 2. Responses and percentile scores in the MC test (N = 122)

	Congruent collocations				Incongruent collocations			
Word class	Correct	%	Incorrect	%	Correct	%	Incorrect	%
AN[a]	929	76.15%	291	23.85%	792	64.92%	428	35.08%
VN[a]	913	74.84%	307	25.16%	765	62.70%	455	37.30%
Total[b]	1842	75.49%	598	24.51%	1557	63.81%	883	36.19%

[a]Max score = 1,220 ($N = 122 \times K = 10$). [b]Max score = 2,440 ($N = 122 \times K = 20$).

Table 3. Summary of the best-fit mixed logit model for MC scores (N = 4,880, log likelihood = –2,334.10)

	β	*SE*	Wald *Z*	*p*	Exp(β)
(Intercept)	1.63	0.27	6.01	<.001**	5.09
VLT score	0.11	0.01	12.44	<.001**	1.12
Item type: Incongruent	–0.79	0.37	–2.14	.03*	0.45

The model has random intercepts for participants and items. *$p < .05$. **$p < .001$.Formula:
glmer (*MCQScore ~ VLTScore + ItemType* + [*1*|*ParticipantCode*] + [*1*|*ItemCode*], *family = "binomial"*)

VLT scores seems to have a large effect with 12% more chances of a correct answer as the VLT score increased by 1 point. The effect of congruency seems to be even bigger with 2 times higher odds of a correct answer for congruent than incongruent collocations. Finally, none of the item-related variables (including word class) contributed significantly to the model, neither did the interaction between VLT and item type.

Thus, to answer the research questions, congruency had an overall effect on untimed collocation recognition. Estimated L2 proficiency also had a sizable effect with higher collocation scores as VLT scores increased. Finally, lack of interaction between the two main predictors seems to suggest that the effect of L1 congruency is omnipresent across the board regardless of the estimated proficiency level.

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

Experiment 2

This experiment uses the same target items but employs a different collocation measure, timed acceptability judgment task, which is assumed to tap into automatic processing. The participants are recruited from the same L1 Arabic–L2 English population.

Methods

Participants

One hundred and six female Arab learners of English took part in this experiment. None of them took part in Experiment 1, but they were sampled from BA students in the same university in Saudi at two levels: preparatory-year program ($N = 55$) and seniors ($N = 51$). They ranged in age between 17 and 25 ($M = 20.10$, $SD = 1.90$). None of them had lived in an English-speaking country for more than 3 months, and they started learning English at school age ($M = 6.74$ years, $SD = 3.66$).

The average scores obtained by the two groups on the 1k and 2k levels of the Updated VLT were 28.65 ($SD = 2.01$) and 24.58 ($SD = 4.82$), respectively. The total VLT average score achieved across the two levels was 53.24 ($SD = 6.52$). The Kuder–Richardson (Formula 21) reliability coefficient for the total VLT score was found acceptable (0.87). Figure 2 presents the distribution of scores. Like Experiment 1, the VLT scores were centred and included in the mixed-effects model as an estimated continuous measure of proficiency.

In addition to the two groups of nonnatives, a group of 37 native speakers of English (female = 21, male= 16) were also included in this experiment to set a baseline for nativelike timed recognition. They ranged in age between 24 and 68 ($M = 45.51$, $SD = 10.73$).

Items

Target collocations for the acceptability judgment task were the same 40 items tested in Experiment 1. In addition, as a baseline for comparison, each target item was matched with a control noncollocate pair that included the same noun but a different verb or adjective. Some of these verbs or adjectives came from distractors in the MC (see Experiment 1). For example, **narrow ability* and **attract a cold* were included as control pairs for *limited ability* and *catch a cold*, respectively. Modifiers in VN collocations are the same used in Experiment 1. All adjectives and verbs in control noncollocate pairs were highly frequent, and no noun, verb, or adjective was repeated more than once in the list. The COCA frequency of control pairs ranged between 0 and 8 with a negative MI score signaling no association. They were also semantically plausible to avoid any inherent advantage for target over control (a limitation of previous research, see the Background section above). Similar to Experiment 1, all frequencies were extracted from the COCA and were log transformed.[6] The full list of target items and control pairs are presented in Appendix S.2 in the online supplemental material. Similar to Experiment 1, item-related variables were statistically controlled for in the mixed-effects model as will be explained in the Analysis section below.

In order to check the difference in corpus-derived frequency between target collocations and control pairs, an analysis of variance with post hoc Tukey comparisons

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

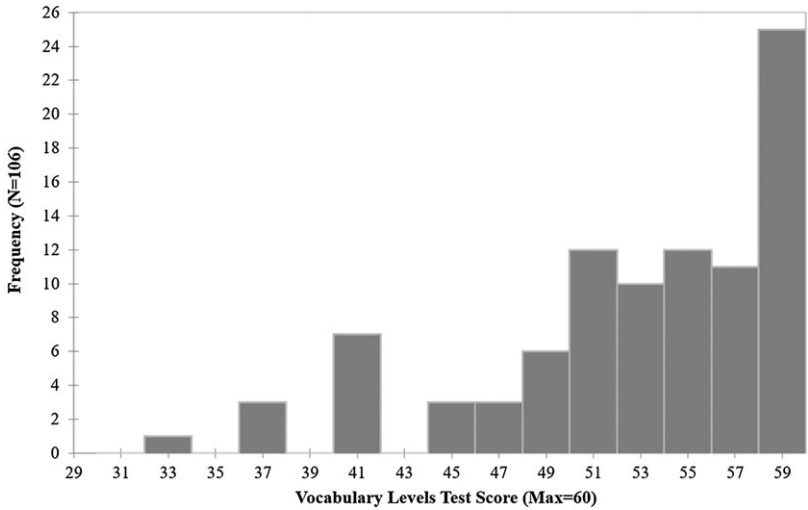

Figure 2. Distribution of VLT scores achieved by participants (Experiment 2).

was conducted. Results showed a significant difference in log collocation frequency between control items ($M = 0.26$, $SD = 0.32$) and congruent collocations ($M = 2.65$, $SD = 0.51$, $d = 2.40$, $p < .001$) as well as incongruent collocations ($M = 2.39$, $SD = 0.38$, $d = 2.13$, $p < .001$), but no difference between the two types of collocations ($d = 0.26$, $p = .09$).

In addition, a norming study was conducted where 18 native speakers of English who did not take part in the main experiment were presented with target collocations and control pairs in two counter-balanced lists and were instructed to rate them for familiarity on a scale from 1 to 7 (1= *never heard the phrase before*, 7 = *heard it many times*). This step was intended to confirm the frequency-based division of items into target collocations and control pairs. An analysis of variance confirmed the corpus-based division of items with a significant difference between control items ($M = 2.67$, $SD = 0.74$) and both congruent ($M = 6.18$, $SD = 0.57$, $d = 3.51$, $p < .001$) and incongruent collocations ($M = 6.13$, $SD = 0.51$, $d = 3.63$, $p < .001$); but no difference between the true collocations ($d = 0.12$, $p = .82$).

Measures

The measure employed in this experiment is a timed phrase acceptability judgment task. As evident from the review of previous research, this is the most common type of task used in collocation research to gauge timed/automatic recognition. The task was designed and administered using E-Prime 3.00 (Psychology Software Tools Inc.; https://www.pstnet.com). Participants had to decide whether each presented pair is commonly used in English using YES/NO-labelled buttons on a gamepad (IHK LLC, Shenzhen China; https://h3-grup.com). Research suggests that gamepads are more sensitive in recording RT than keyboards (see Segalowitz & Graves, 1990; Shimizu, 2002).

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

The 40 control and 40 target items were counterbalanced into two lists (Task 1 and Task 2) so that no control/target appeared in the same list more than once. Each list included 10 AN collocations, 10 VN collocations, 10 AN control pairs, and 10 VN control pairs. In addition to these, 42 filler collocate and noncollocate pairs were included in the list (21 AN and 21 VN) to make the task as natural as possible. None of these fillers included any component word from the target collocations or control pairs. The list was divided into two parts with a break in the middle based on word class. In Task 1, the AN section preceded the one including VN pairs while in Task 2 the order was reversed. This was intended to control for the effect of word class presentation order on RT and accuracy scores.

There were 10 practice trials in the beginning to familiarize participants with the task. Thus, in each task, participants were presented with 92 pairs in total (10 practice, 20 targets, 20 controls, and 42 fillers). Half of the participants completed Task 1 (natives $N = 17$, nonnatives: $N = 55$) and the other half completed Task 2 (natives: $N = 20$, nonnatives: $N = 51$). Each trial started with a fixation screen of 12 asterisks, after 1 s, it was replaced by a blank screen for 50 ms, finally the pair appeared and stayed on the screen for 5000 ms or until a response was detected. The order of trials was randomized across subjects, and E-Prime recorded accuracy and response latency for each trial.

Procedures

The E-Prime task was conducted in a quiet computer lab and lasted 15 min. The instructions were adapted from Wolter and Gyllstad (2013, p. 460) and read as follows:

> You need to decide whether the combinations presented are commonly used in English or not. Press the YES button (labelled "Yes" on the gamepad) if the word combination is commonly used in English. Press the NO button (labelled "NO" on the gamepad) if the word combination is NOT commonly used in English. Please answer as accurately and quickly as possible.

After completing the acceptability judgment task, the nonnatives were administered the two levels of the Updated VLT and completed them in 15 min. Finally, all participants completed a language background questionnaire.

Analysis

The analysis was conducted on R version 3.5.0. (R Core Team, 2018). Accuracy scores (0/1) for natives and nonnatives were analyzed separately in two mixed-logit models for binary data (*glmer* function in the *lme4* package). Then, we analyzed predictors of our main dependent variable in this experiment, RT data, through two linear mixed-effects (LME) models (*lmer* function in the *lme4* package), one for natives and the other for nonnatives. The analysis for natives was conducted to set a baseline against which the nonnative recognition can be compared.

Initially, timed-out responses were removed, and the data was checked for outliers. Responses that were faster than 500 ms were excluded from further analysis. As for the highest cutoff point, 2000 was set for natives and 4000 was set for non-natives.[7]

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

This resulted in excluding 3.58% and 1.36% of data points for natives and nonnatives, respectively.

In the mixed-logit analysis for accuracy, a response was coded either as 1 (*correct*) or as 0 (*incorrect*) based on whether it matched the expected response: 1 (Yes for a collocation and No for a control pair) and 0 (No for a collocation and Yes for a control pair). The models included accuracy binary scores as the dependent variable and participant/items as random variables. The two main fixed effects were then tested: item type (congruent, incongruent, or control) and estimated proficiency (VLT scores for nonnatives only). In addition to these, the models also included the interaction between the two main factors, trial number, and five item-related variables: word class (AN vs. VN), word 1 log frequency, word 2 log frequency, log pair frequency, and pair length. Effect size was estimated based on odds ratios (Exp[β]) transformed from log odds.

We then explored variables predicting RT in LME models. Only correct responses were included in this analysis with RT as the dependent measure (after log transformation), participants/items as random variables, and the same fixed effects tested above. Effect sizes for these models are represented by marginal and conditional R^2 values. The former involves only fixed effects while the latter incorporates random effects as well (see Wolter & Yamashita, 2018, for an explanation).

Collinearity among item-related variables was checked prior to the analysis and residuals were calculated for word 1 frequency, word 2 frequency, pair frequency, and pair length. All residuals correlated well with the original variables (log word 1 frequency $r = .85$, $p < .001$; log word 2 frequency $r = .74$, $p < .001$; pair length $r = .94$, $p < .001$, log pair frequency $r = .44$, $p < .001$). Table 4 below presents a summary of continuous variables.

Results and discussion of Experiment 2

Table 5 below presents RTs and accuracy scores in the acceptability judgment task for both natives and nonnatives. In general, like earlier research, natives were faster and more accurate than nonnatives in making decisions (but see Gyllstad & Wolter, 2016, for an exception).

RT data seems to suggest an evident processing lag for control pairs in comparison to target collocations with an average difference of around 215 ms and 400 ms for natives and nonnatives, respectively. Comparing congruent and incongruent collocations, in contrast, does not seem to manifest a clear trend in RT for either group. The same holds true for word class differences. However, it should be noted that the RTs presented in Table 5 represent raw values before controlling for estimated proficiency and for item-related variables.

Looking now at accuracy scores, it seems that natives processed both types of collocations similarly (with a range of 95% and 98% accurate scores) but were less accurate in making decisions about control pairs (around 80% correct responses). Conversely, nonnatives showed a clear gradual increase in the percentage of correct answers from control items with around 50%, to incongruent collocations with over 70%, and finally congruent collocations with above 86%.[8] These percentages do not seem to be affected by word class with fairly similar scores for AN and VN collocations.

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

Table 4. Summary of continuous variables (acceptability judgment task)

Variable	Range (adjusted range)	*SD*	*Mdn*
Trial number	1–41 (–20.12 – 19.88)	11.84	–0.12
Resid log pair frequency	0–8.18 (–1.87 – 2.06 log units)	0.89	–0.16
Resid pair length	9.00–23.00 (–5.58 – 7.84 characters)	2.88	–0.24
Resid word 1 frequency	8.43–13.14 (–2.23 – 1.83 log units)	0.77	0.06
Resid word 2 frequency	9.19–13.72 (–2.35 – 1.42 log units)	0.74	0.07
VLT score (nonnatives only)	32.00–60.00 (–21.97 – 6.03 points)	6.44	2.03

Note: The second column shows the range of the variables. The adjusted range after transformation and/or centring, is presented in parentheses. Standard deviations and medians refer to the predictor values in the models. All variables are centred, and their means are zero.

Table 5. Mean RT (standard error; SE) and accuracy scores (percentage) in the acceptability judgment task

		RT (*SE*)			Accuracy score (%)		
Group	Word class	Congruent	Incongruent	Control	Congruent	Incongruent	Control
Natives	AN	891.87	875.16	1109	179	176	269
$N = 37$		(18.43)	(20.81)	(18.19)	(98.35%)	(97.78%)	(77.52%)
	VN	978.14	914.81	1151.74	177	171	291
		(22.4)	(17.74)	(17.05)	(95.68%)	(95.00%)	(83.86%)
Nonnatives	AN	1525.29	1436.25	1893.99	466	417	536
$N = 106$		(25.95)	(27.74)	(29.52)	(89.10%)	(78.68%)	(52.09%)
	VN	1528.97	1588.34	1971.4	446	388	592
		(29.28)	(30.17)	(26.52)	(85.93%)	(73.90%)	(57.14%)

Natives

Accuracy. Natives' best-fit mixed-logit model for accuracy scores is presented in Table 6. Only item type was found significant with almost 11 times more chances of a correct decision for both congruent and incongruent collocations in comparison to control items. In order to explore the difference between congruent and incongruent collocations, we fit the same model after setting "incongruent" as the reference level. Results showed that, as expected, natives did not differ in their decision accuracy for both types of collocations (Wald $Z = 0.02$, $p = .98$, $\text{Exp}[\beta] = 1.01$).

Response latency. Table 7 presents the best-fit model for variables predicting natives' response latencies. The first significant effect is corpus-derived pair frequency with shorter RT (faster processing) as the frequency increased. Then, pair length was significant with longer RT (slower processing) as the length increased. Finally, item type was significant with faster processing for both types of collocations in comparison to

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

Table 6. Summary of the best-fit mixed logit model for natives' accuracy scores (N = 1,421, log likelihood = −406.8)

	β	*SE*	Wald *Z*	*p*	Exp(β)
(Intercept)	1.83	0.25	7.27	<.001*	6.26
Item type: Congruent	2.38	0.45	5.33	<.001*	10.77
Incongruent	2.36	0.45	5.20	<.001*	10.63

The model has random intercepts for participants and items. *$p < .001$. Formula:
glmer (*JudgmentAccuracy ~ ItemType + [1|ParticipantCode) + [1|ItemCode], family = "binomial"*)

Table 7. Summary of the best-fit LME model for variables predicting natives' RT data (N = 1,263, R^2marginal = 0.20, R^2conditional = 0.57)

	Estimate	*SE*	*t* value	Pr(>|*t*|)
(Intercept)	7.02	0.03	229.54	<.001**
Resid log pair frequency	−0.03	0.01	−2.89	.01*
Resid pair length	0.02	0.00	4.81	<.001**
Item type: Congruent	−0.23	0.02	−9.35	<.001**
Incongruent	−0.25	0.02	−10.11	<.001**

The model has random intercepts for participants and items. *$p < .05$. **$p < .001$. Formula:
lmer(*LogRT ~ ResidLogPairFrequency + ResidPairLength + ItemType + [1|ParticipantCode] + [1|ItemCode*])

noncollocate pairs. Redefining the reference level as "incongruent" showed no difference between natives' processing of congruent and incongruent collocations (t value = 0.74, $p = .46$). This model explains almost 60% of variance in the data as suggested by the conditional R^2 value.

To summarize results from natives, the analysis showed a clear advantage in both decision accuracy and processing time for target items over control pairs but almost identical processing for both types of collocations. These findings confirm our previous norming rating results (see the Items section above), goes in line with previous research findings (e.g., Wolter & Gyllstad, 2013), and sets a baseline for comparison with the main nonnative group.

Nonnatives

Accuracy. Table 8 below presents the best-fit model for variables predicting accuracy scores. VLT scores and item type contributed significantly to the model with no interaction between the two. The odds of a correct answer in the acceptability judgment task was 6% bigger as the VLT increased by one unit. As for congruency, both congruent and incongruent collocations led to more correct answers in comparison to the control reference level (almost 8 and three times bigger, respectively). Finally, a model where "incongruent" was the reference level revealed significantly more correct responses for congruent than incongruent collocations (Wald $Z = 3.30$, $p < .001$, Exp[β] = 2.55).

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

Table 8. Summary of the best-fit mixed logit model for non-natives' accuracy scores (N = 4,162, log likelihood = –2,220.2)

	β	*SE*	Wald *Z*	*p*	Exp(β)
(Intercept)	0.27	0.14	1.94	.05	1.30
VLT score	0.06	0.01	7.48	<.001*	1.06
Item type: Congruent	2.05	0.25	8.28	<.001*	7.77
Incongruent	1.11	0.24	4.72	<.001*	3.04

The model has random intercepts for participants and items. *$p < .001$. Formula:
glmer(JudgmentAccuracy ~ VLTScore + ItemType + [1|ParticipantCode] + [1|ItemCode], family = "binomial")

Response latency. The best-fit LME model for nonnatives' RT data is presented in Table 9 below. As indicated by the conditional R^2 value of the model, it explains 47% of variance in the data. trial number, log pair frequency, and pair length had significant effects with longer RTs as the experiment went on (showing a fatigue effect for nonnatives) and as the length of the pair increased, but shorter RTs as the frequency increased. Word class and the frequency of individual words, however, did not add any significant value to the model.

As for the main fixed factors, item type had a main effect with slower responses for control items in comparison to both congruent and incongruent collocations. Upon redefining the refence level as "incongruent," we found that the two types of collocations did not differ in RT in the final model (t value = –0.46, $p = .65$). Finally, VLT scores did not show up as a significant main effect but rather significantly interacted with item type.

Figure 3 depicts the interaction between the two variables along with the effect size for each item type revealing quite interesting results. Looking at RTs to the three types of items (congruent, incongruent, and control) and how they changed as the VLT scores increased, one can note the following points. First, RTs for all items decreased as estimated proficiency increased. Second, nonnatives achieving low scores in the VLT seem to process congruent collocations normally with a big RT advantage over noncollocate pairs. However, incongruent pairs do not seem to have gained their collocation status yet as their processing is rather similar to baseline pairs. Third, while both congruent and baseline items seem to show a gradual decrease in processing time, incongruent collocations demonstrate sharp and substantial changes (with the biggest effect size). This led to more nativelike processing (see above) as estimated proficiency increased, that is, smaller differences between congruent and incongruent collocations and larger differences between them and control pairs.[9] This finding is further supported by the statistical figures revealed upon redefining the reference level for item type to "incongruent." In that model, the interaction with VLT scores was still significant ($t = 2.60$, $p = .009$). This stands in sharp contrast with the lack of variance between congruent and incongruent collocations (as a main effect) reported above. Thus, one might conclude that there was not an inherent difference between congruent and incongruent collocations. Rather, they only showed a difference in recognition time as estimated proficiency (VLT scores) decreased.

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

Table 9. Summary of the best-fit LME model for variables predicting non-natives' RT data (N = 2,845, R^2marginal = 0.19, R^2conditional = 0.47)

	Estimate	*SE*	*t* value	Pr(>|*t*|)
(Intercept)	7.53	0.02	321.33	<.001***
Trial	0.00	0.00	−2.40	.02*
Resid log pair frequency	−0.04	0.01	−3.03	.003**
Resid log pair length	0.02	0.00	6.09	<.001***
VLT score	0.00	0.00	−1.53	.13
Item type: Congruent	−0.27	0.03	−10.37	<.001***
Incongruent	−0.26	0.03	−9.70	<.001***
VLT Score × Item Type: Congruent	−0.01	0.00	−2.64	.008**
VLT Score × Item Type: Incongruent	−0.01	0.00	−5.16	<.001***

The model has random intercepts for participants and items. *$p < .05$. **$p < .01$. ***$p < .001$. Formula: *lmer*(*LogRT ~ Trial + ResidLogPairFrequency + ResidPairLength + VLTScore*ItemType + [1|ParticipantCode] + [1|ItemCode]*).

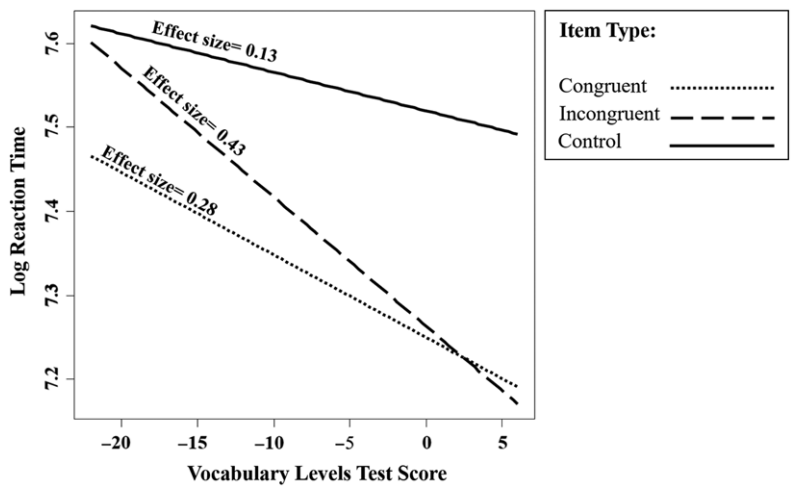

Figure 3. Interaction between VLT score (centred) and item type for the nonnative RT data (Experiment 2).

Thus, unlike untimed recognition (see Experiment 1) where congruency and proficiency effects are omnipresent, timed processing manifests a proficiency-modulated advantage for congruent over incongruent collocations. The L1 effect seems to be strong for lower level EFL learners, but it fades away gradually as estimated proficiency increases approaching nativelike collocation processing.

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

General discussion

This study aimed at dealing with limitations of previous research, which investigated the effect of congruency and proficiency on nonnative collocation processing and the interaction between the two factors. To that end, we conducted two experiments with Arab EFL learners, one evaluating untimed collocation recognition in a traditional MC test and the other assessing timed recognition in an acceptability judgment task.

Results showed that congruency had a substantial effect on untimed collocation recognition with more correct answers for congruent than incongruent collocations across proficiency levels in the MC test (Experiment 1). These findings are in line with previous research employing MC (Nguyen & Webb, 2017) and Yes/No untimed measures (Wolter & Gyllstad, 2011). An interesting point to note here is that our participants scored higher in the MC test (with a rough average of 70%) than Nguyen and Webb's (2017) Vietnamese EFL learners who knew only 50% of the target items. This difference in gains might be attributed to the fact that collocations in the present study included those comprising the most frequent 2,000 word families in English while Nguyen and Webb's items expanded the selection to collocates at the 3,000 word-family level.

Turning now to timed collocation processing, RT data in Experiment 2 (see Figure 3) seems to suggest that the L1 congruency influence is closely tied to proficiency level with smaller effect as estimated proficiency increases. These results do not only go in line with Yamashita and Jiang's (2010) findings in the Japanese EFL context but also add to them through demonstrating a gradual fading away of the L1 effect. However, they stand in sharp contrast with Wolter and Gyllstad (2013) and Wolter and Yamashita (2018) who did not show any interaction between proficiency and congruency during real-time recognition. As indicated in the Background section above, an important limitation of these two studies is the fact that they either (a) dealt with proficiency as a categorical variable (Wolter & Yamashita, 2018), or (b) tested participants at one level of proficiency (Wolter & Gyllstad, 2013), which may have veiled such effects. One point to note here is that it is not possible to directly compare the proficiency level of our participants to those in previous research employing the same timed task (e.g., Wolter & Gyllstad, 2013; Wolter & Yamashita, 2018; Yamashita & Jiang, 2010) due to variation in proficiency measures. However, although our estimated proficiency measure was also fairly limited (see the Limitations section below), including proficiency as a continuous rather than a categorical variable seems to have allowed for a clearer depiction of its effect on collocation recognition.

Other determinants of L2 collocation recognition

Although our study focused on two main factors of L2 collocation recognition (i.e., congruency and proficiency), previous research has pointed out to several other important effects including word class, collocation frequency, and the frequency of individual component words. These were included in our mixed-effects models and showed interesting findings in both experiments.

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

On the untimed side of processing, in contrast with Nguyen and Webb (2017) where word and collocation frequency were significant determinants of MC test scores, Experiment 1 in the present study did not find such effects. This might be due to the fact that Nguyen and Webb experimentally manipulated the frequency of individual words and collocations (at three distinct levels) allowing such factors to show up. Our study did not specifically manipulate frequency either at the word or at the collocation level leading to lack of variation. As for word class (VN vs. AN), earlier corpus-based studies (Laufer & Waldman, 2011; Nesselhauf, 2003) reported higher difficulty in the production of VN than AN collocations. Our results, however, agree with those reported by Nguyen and Webb (2017), showing no significant difference between the two types of collocations. It seems that nonnatives process them similarly, at least at the recognition level.

Looking now at the acceptability judgment task, our results exhibited the collocation-frequency effects found in previous studies that gauged timed processing (e.g., Wolter & Gyllstad, 2013; Wolter & Yamashita, 2018; see also Sonbul, 2015). This finding supports usage-based models of language development, which claim that experience with the language is a major predictor of language processing and development (Bybee, 2006) both in the L1 and L2. However, our findings go against Wray's (2005) claim that natives and nonnatives process formulaic language differently. Experiment 2 results seem to show that as estimated proficiency increases, nonnatives process collocations on par with natives (see Gyllstad & Wolter, 2016, for similar findings). Along the same lines, Groom (2009) showed that, with more immersion in the L2 context, nonnatives seem to exhibit more nativelike collocation patterns in written production.

When it comes to word-level frequency effects, though, insignificant effects in the present study stand in contrast with Wolter and Yamashita (2018). As reported above, this might be related to a ceiling effect as all component words in the collocations were highly frequent. Finally, as for word class effects on real-time processing, our results show that, like Wolter and Yamashita (2015), nonnatives do not differ in their timed processing of AN and VN collocations. This result is interesting in our context as the word order in Arabic is the same for VN collocations but is the opposite for AN collocations. Despite this difference in word order, it seems that Saudi EFL learners in the present study processed them similarly. This seems be due to the long experience, an average of about 13 years (from 6.74 years to 20.10 years, see the Methods section in Experiment 2), they had with English. This might have allowed them to overcome this mismatch in word order between the two languages. It would be interesting to explore the effect of word class configuration with beginner Arab EFL learners.

Theoretical implications

Although the current study did not originally aim to address various models of bilingual language processing, it seems to have certain tentative implications. Previous research employing timed collocation measures (e.g., Wolter & Gyllstad, 2013; Yamashita & Jiang, 2010, see the Background section) has attempted to extend theoretical models of bilingual word processing to collocations. Yamashita and Jiang (2010), for example, took their findings as support of the revised hierarchical model (RHM; Kroll & Stewart, 1994), which assumes a nonselective view of processing in the early stages of L2 learning

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

whereby both languages are activated making it difficult to suppress one over the other (i.e., L2 is accessed through the L1). As L2 proficiency increases, a more direct, faster route to the L2 lexicon is possible.

The present study adds to Yamashita and Jiang's (2010) findings through (a) including EFL participants only, thus controlling for L2 context, and (b) treating proficiency as a continuum to allow for a more fine-grained depiction of its effect. Our results seem to lend support to the RHM as the nonnatives with low estimated levels of proficiency showed dependence on the L1, but the ones at higher levels seemed to follow a more direct route to lexical access exhibiting more nativelike processing (see Figure 3 above).

Thus, despite recent calls for leaving the RHM behind due to evidence that it fails to account for various aspects of L2 lexical processing (see, for example, Brysbaert & Duyck, 2010), our results seem to suggest that predictions of the RHM still hold value. It should be noted, however, that results of the present study should be treated rather cautiously as proficiency was only estimated based on a short vocabulary test rather than a standardized measure. In addition, our results stop at the level of lexical access/processing. Models of bilingual lexical processing also make assumptions about knowledge representation (see Kroll & Tokowicz, 2005 for an overview), but this was not the focus of the present study.

In contrast with the RHM, proponents of the age of acquisition/order of acquisition effects on bilingual lexical processing claim that the most important determinant of word recognition (both time and accuracy) is the order in which the words were learned (see, for example, Izura et al., 2011). Wolter and Gyllstad's (2013) and Wolter and Yamashita's (2018) findings are interpreted within these effects whereby congruent collocations, which are plausible in the L1, are learned first and will always show a processing advantage over incongruent collocations. However, our results seem to suggest that this explanation might not be precise and might be concealed by the narrow range of proficiency in these earlier studies. Our findings seem to suggest that, unlike frequency effects, which uphold even at the highest levels of L2 proficiency, congruency effects seem to fade away as proficiency increases.

Another similar explanation of the congruency effect is Jiang's (2000) model of L2 lexical fossilization (see Wolter & Gyllstad, 2011). Jiang divides lexical knowledge into two levels: lexeme (phonology, orthography, and morphology) and lemma (syntax and semantics), claiming that the latter is amenable to influence from the L1 even at very advanced L2 levels and might result in a fossilized L2 state. The model also claims that only exceptional nonnative cases might be able to escape these L1-lemma effects to achieve nativelike performance. Looking at Figure 3 and the systematic trend for more nativelike recognition as estimated proficiency increases, one might possibly refute such claims of resistant L1 effect on L2 processing. Our results seem to suggest that with enough L2 exposure, nonnative learners might most likely be able to overcome influences from the L1, achieving more nativelike recognition (cf. Wolter & Gyllstad, 2011, 2013).

Pedagogical and research implications

The study has important implications for L2 teaching, testing, and research practice. Since the L1 effect seems to diminish gradually as proficiency increases, L2 learners

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

can be pushed out of the fossilized state of lexical development (Jiang, 2000; see above) through more exposure to language. This can eventually lead to dealing with the L2 input through a nativelike eye removing the L1 lenses that might overshadow processing in the early stages of L2 learning. However, the question remains: how best can L2 learners be provided with such an exposure especially in the EFL context where contact with English is often limited to classroom time? One feasible option is extensive reading programs, which, as argued by Nation (2011), have been underapplied in EFL classrooms in spite of their well-documented value. These programs can expand vocabulary knowledge and help learners develop lexical aspects that are not usually amenable to teaching such as collocations. This is not to say that collocations should not be taught explicitly. Previous research has shown direct collocation teaching as an effective tool in developing knowledge collocations (e.g., Laufer & Girsai, 2008), but teachers need to be careful about which collocations to teach and how best to teach them.

Although the gains obtained from Experiment 1 and Experiment 2 might not be directly comparable (correct/incorrect responses in Experiment 1 vs. RT for correct responses only in Experiment 2), some broad implications related to L2 collocation testing can be drawn. Previous research has mostly used untimed paper-and-pencil tests and has made claims regarding congruency (e.g., Bahns & Eldaw, 1993; Biskup, 1992) and the way it interacts with learning conditions (e.g., Peters, 2016). Our results showed that congruency has differential effects on untimed and timed L2 collocation processing. It seems that nonnative users always exhibit L1 effects when they are prompted to answer untimed tests. However, during time-pressured processing, the L1 effect seems to disappear as estimated proficiency increases. Based on these findings, one might tentatively propose that language teachers need to interpret results of paper-and-pencil tests cautiously as they do not necessarily translate into automatic processing (see Sonbul & Schmitt, 2013, for similar claims). Moreover, research on L2 collocation learning and processing should start to incorporate both types of measures in order to capture congruency effects more closely. To the best of our knowledge, only one study (Szudarski & Conklin, 2014) attempted to explore the role that L1 plays in L2 collocation learning using a timed acceptability judgment task.

One last implication of the present study is that the semantic deviation of control items in previous research (see the Background section above) does not necessarily refute earlier findings on collocation processing. Even when control items were made semantically plausible, both natives and nonnatives processed them more slowly than real collocations.

Limitations and suggestions for future research

Our study is limited in several ways. First, English language proficiency was assessed using knowledge of the most frequent 2,000 word families only. Including higher levels might have resulted in expanding the variation in proficiency levels, leading to cleaner depiction of the interaction with congruency. The study also did not explore the effect of transparency which has been shown to highly correlate with congruency (see Gyllstad & Wolter, 2016; Yamashita, 2018).

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

Second, we only tested recognition of two types of collocations (VN and AN). A recent study by Siyanova-Chanturia and Janssen (2018) employed a timed recall measure (phrase elicitation task) in exploring the processing of formulaic language (i.e., binomials). Future research on determinants of L2 collocation processing and/or leaning may need to start assessing recall using such psycholinguistic techniques to tap into spontaneous productive language use. Moreover, findings for VN and AN collocations might need to be expanded through investigating other configurations (e.g., verb + adverb, verb + preposition, adverb + adjective) as the learning/processing burden may not be the same for all word classes.

A third limitation of the present study is related to the nature of the tasks used. Our timed task required an explicit judgment. Although such a task has been common in previous research, other psycholinguistic measures can more directly tap into natural language processing (e.g., self-paced reading, primed LDT, eye tracking, and event-related potentials). Although several studies have explored frequency effects on formulaic language using such sensitive tasks (see, for example, Siyanova-Chanturia, Conklin, & van Heuven, 2011; Sonbul, 2015), very few studies (Carrol & Conklin, 2014, 2017; Carrol et al., 2016; Wolter & Gyllstad, 2011) have explored congruency effects on L2 processing of formulaic language online (i.e., in real time). Employing such sensitive measures might lead to differential effects of congruency and proficiency and the interaction between them. Turning now to the untimed MC test, it is limited in that it only gauged collocation form without considering meaning. Nguyen and Webb (2017) suggested that future research may need to focus on predictors of meaning recall/recognition.

Fourth and finally, we were only able to recruit female learners. Future research may need to include male learners as well to evaluate the generalizability of results to both male and female Arab EFL learners.

Conclusion

This study aimed at exploring the effect of L1 on L2 collocation recognition and how this might be influenced by estimated levels of proficiency. In two experiments, we assessed both untimed and timed L2 collocation recognition in the underresearched Arabic EFL context. Results showed that the L1 influenced the ability to answer MC items regardless of proficiency level while the effect on timed recognition faded away as L2 competence increased. These results go in line with Yamashita and Jiang's (2010) findings and point out to an important distinction to be considered in future collocation research between untimed collocation processing and the ability to use collocations fluently in real time.

Acknowledgments. The researchers thank Prince Sultan University for funding this research project through the research lab (Applied Linguistics Research Lab Grant RL-CH-2019/9/1/0). We would also like to thank four anonymous reviewers for their useful comments on an earlier draft of this article that helped us to improve it significantly.

Supplementary material. To view supplementary material for this article, please visit: https://doi.org/10.1017/S014271642000051X.

Notes.

1. It should be noted that the present study stops at the level of processing and does not make any assumptions about explicit/ implicit knowledge representation.
2. We could not calculate the Cronbach alpha coefficient for the VLT since we only had access to the total score per level, not the raw 0/1 scores. Thus, we opted to use the Kuder–Richardson (Formula 21) coefficient in both experiments as a rough estimate of reliability.
3. The study (both experiments) was approved by the Research Ethics Committee of the Saudi University. Upon receiving clearance, the second author and a research assistant visited classrooms and explained the purpose and requirements of the study and sought the students' consent to participate. The students who gave their informed consent were invited to a quiet lab in designated times. They were made aware that participation in the study was voluntary and would lead to no rewards.
4. For more details on fitting mixed-effects models, see Sonbul and Schmitt (2013).
5. To interpret the effect of a variable in percentage terms, odds ratios are multiplied by 100. When the value is negative, one is divided by the (Exp[β]) value to interpret its effect.
6. A value of 1 was added to all pair frequencies prior to log transformation. This was intended to avoid logging the zero frequency of certain control items into undefined values.
7. We increased the upper cutoff point for the L2 learners as they were obviously slower than natives in processing the presented pairs.
8. It should be noted that the accuracy percentage for nonnatives in the present study is far lower than that reported by Wolter and Gyllstad (2013) who used the same acceptability judgment task: an average of 50% versus 70%, respectively. This might be due to the fact that control items in the present study were semantically plausible making it extremely difficult for EFL learners to make an accurate "NO" decision.
9. In order to check the validity of the 1k and 2k VLT scores as an estimated measure of proficiency, we conducted an additional analysis with self-reported proficiency (subjective ratings obtained from the nonnatives) in place of the VLT scores. Significant predictors in the best-fit model were the same as those reported in Table 9.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alharbi, R. M. S. (2017). *Acquisition of lexical collocations: A corpus-assisted contrastive analysis and translation approach*. Unpublished PhD thesis, Newcastle University.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System*, **21**, 101–114.

Biskup, D. (1992). L1 influence on learners' renderings of English collocations: A Polish/German empirical study. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 85–93). London: Palgrave Macmillan.

Brysbaert, M., & Duyck, W. (2010). Is it time to leave behind the revised hierarchical model of bilingual language processing after fifteen years of service? *Bilingualism: Language and Cognition*, **13**, 359–371.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, **82**, 711–733.

Carrol, G., & Conklin, K. (2014). Getting your wires crossed: Evidence for fast processing of L1 idioms in an L2. *Bilingualism: Language and Cognition*, **17**, 784–797.

Carrol, G., & Conklin, K. (2017). Cross language lexical priming extends to formulaic units: Evidence from eye-tracking suggests that this idea 'has legs.'" *Bilingualism: Language and Cognition*, **20**, 299–317.

Carrol, G., Conklin, K., & Gyllstad, H. (2016). Found in translation: The influence of the L1 on the reading of idioms in a L2. *Studies in Second Language Acquisition*, **38**, 403–443.

Davies, M. (2008) The Corpus of Contemporary American English (COCA): 560 million words, 1990–present. Available online at https://corpus.byu.edu/coca/

https://doi.org/10.1017/S014271642000051X Published online by Cambridge University Press

El-Dakhs, D. A. S. (2015). The lexical collocational competence of Arab undergraduate EFL learners. *International Journal of English Linguistics*, **5**, 60–74.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, **27**, 141–172.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text–Interdisciplinary Journal for the Study of Discourse*, **20**, 29–62.

Farghal, M., & Obeidat, H. (1995). Collocations: A neglected variable in EFL. *International Review of Applied Linguistics in Language Teaching*, **33**, 315–331.

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Language tasks: Teaching, learning and testing* (pp. 75–93). Harlow: Longman.

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.

Godfroid, A., Loewen, S., Jung, S., Park, J. H., Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge: Evidence from eye-movement patterns. *Studies in Second Language Acquisition*, **37**, 269–297.

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.

Groom, N. (2009). Effects of second language immersion on second language collocational development. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language* (pp. 21–33). New York: Palgrave Macmillan.

Gyllstad, H. (2007). *Testing English collocations*. Unpublished doctoral dissertation, Lund University, Lund, Sweden.

Gyllstad, H. (2009). Designing and evaluating tests of receptive collocation knowledge: COLLEX and COLLMATCH. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language* (pp. 153–170). New York: Palgrave Macmillan.

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

Hussein, R. F. (1990). Collocations: The missing link in vocabulary acquisition amongst EFL learners. *Papers and Studies in Contrastive Linguistics*, **26**, 123–136.

Izura, C., Pérez, M. A., Agallou, E., Wright, V. C., Marín, J., Stadthagen-González, H., & Ellis, A. W. (2011). Age/order of acquisition effects and the cumulative learning of foreign words: A word training study. *Journal of Memory and Language*, **64**, 32–58.

Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics*, **21**, 47–77.

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, **33**, 149–174.

Kroll, J. F., & Tokowicz, N. (2005). Models of bilingual representation and processing: Looking back and to the future. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 531–553). New York: Oxford University Press.

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

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, **61**, 647–672.

Migdad, A. A. (2012). *The role of mother tongue in reception and production of collocations by English majors at the Palestinian universities*. Unpublished MA dissertation, Islamic University of Gaza, Palestine.

Nation, I. S. P. (2011). Research into practice: Vocabulary. *Language Teaching*, **44**, 529–539.

Nation, I. S. P. (2012). *The BNC/COCA word family lists. Document bundled with Range Program with BNC/COCA Lists*, 25. Retrieved from: https://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Information-on-the-BNC_COCA-word-family-lists.pdf

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, **24**, 223–242.

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

Oxford collocations dictionary for students of English. (2002). Oxford: Oxford University Press.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–225). London: Longman.

Peters, E. (2016). The learning burden of collocations: The role of interlexical and intralexical factors. *Language Teaching Research*, **20**, 113–138.

Puimège, E., & Peters, E. (2019). Learners' English vocabulary knowledge prior to formal instruction: The role of learner-related and word-related variables. *Language Learning*, **69**, 943–977.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

Segalowitz, S. J., & Graves, R. E. (1990). Suitability of the IBM XT, AT, and PS/2 keyboard, mouse, and game port as response devices in reaction time paradigms. *Behavior Research Methods, Instruments, & Computers*, **22**, 283–289.

Shimizu, H. (2002). Measuring keyboard response delays by comparing keyboard and joystick inputs. *Behavior Research Methods, Instruments, & Computers*, **34**, 250–256.

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **37**, 776–784.

Siyanova-Chanturia, A., & Janssen, N. (2018). Production of familiar phrases: Frequency effects in native speakers and second language learners. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **44**, 2009–2018.

Sonbul, S. (2015). Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/online collocational processing. *Bilingualism: Language & Cognition*, **18**, 419–437.

Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: Acquisition of collocations under different input conditions. *Language Learning*, **63**, 121–159.

Sonbul, S., & Siyanova-Chanturia, A. (in press). Research on the on-line processing of collocation: Replication of Wolter and Gyllstad (2011) and Millar (2011). *Language Teaching*. Advance online publication. doi: 10.1017/S0261444819000132

Szudarski, P., & Conklin, K. (2014). Short- and long-term effects of rote rehearsal on ESL learners' processing of L2 collocations. *TESOL Quarterly*, **48**, 833–842.

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *International Journal of Applied Linguistics*, **168**, 34–70.

Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics*, **32**, 430–449.

Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing: A comparison of congruent and incongruent collocations. *Studies in Second Language Acquisition*, **35**, 451–482.

Wolter, B., & Yamashita, J. (2015). Processing collocations in a second language: A case of first language activation? *Applied Psycholinguistics*, **36**, 1193–1221.

Wolter, B. & Yamashita, J. (2018). Word frequency, collocational frequency, L1 congruency, and proficiency in L2 collocational processing: What accounts for L2 performance? *Studies on Second Language Acquisition*, **40**, 395–416.

Wray, A. (2005). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Yamashita, J. (2018). Possibility of semantic involvement in the L1-L2 congruency effect in the processing of L2 collocations. *Journal of Second Language Studies*, **1**, 60–78.

Yamashita, J., & Jiang, N. (2010). L1 influence on the acquisition of L2 collocations. *TESOL Quarterly*, **44**, 647–668.

Zhang, X. (2017). Effects of receptive-productive integration tasks and prior knowledge of component words on L2 collocation development. *System*, **66**, 156–167.

Cite this article: Sonbul, S. and El-Dakhs, D. (2020). Timed versus untimed recognition of L2 collocations: Does estimated proficiency modulate congruency effects? *Applied Psycholinguistics* **41**, 1197–1222. https://doi.org/10.1017/S014271642000051X