*Applied Psycholinguistics* (2019), **40**, 987–1017
doi:10.1017/S0142716419000109

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

# Frequency effects on first and second language compositional phrase comprehension and production

Sarut Supasiraprapa*

National Institute of Development Administration, Thailand
*Corresponding author. Email: sarut.sup@nida.ac.th

(Received 04 May 2018; revised 17 January 2019; accepted 22 January 2019)

**Abstract**

Usage-based approaches to language acquisition posit that first (L1) and second language (L2) speakers should process more frequent compositional phrases, which have a meaning derivable from word parts, faster than less frequent ones (e.g., Bybee, 2010; Ellis, 2011). Although this prediction has received increasing empirical support, methodological limitations in previous relevant studies include a lack of control of frequencies of subparts of target phrases and scant attention to L2 production. Addressing these limitations, the current study tested phrase frequency effects in both language comprehension and production in two respective experiments, in which adult native English speakers ($N = 51$) and English L2 learners ($N = 52$) completed a timed phrasal decision task and an elicited oral production task. Experiment 1 revealed phrase frequency effects in both groups, lending support to usage-based researchers' proposal that L1 and L2 speakers retain memory of word co-occurrences and that compositional phrase processing reflects an accumulation of statistics in previously encountered input. Experiment 2, however, provided weaker evidence for phrase frequency effects in these participant groups. Based on the results and previous empirical studies, methodological issues that may have impacted frequency effects and implications for future work in this area are discussed.

**Keywords:** comprehension; frequency effects; L2 learners; production; usage-based

## First language (L1) acquisition from a usage-based perspective

A much-debated issue in L1 acquisition pertains to how compositional phrases (those with a meaning derivable from word parts; e.g., *don't have to worry*, as opposed to sequences like *He left the US for good*, where *for good* cannot be taken apart to derive its meaning) are acquired and processed. The mainstream generative approach to language (Chomsky, 1995; Pinker, 1999; Prasada & Pinker, 1993) argues for a human innate language-specific acquisition device and a separation of the lexicon and the grammar in a speaker's linguistic representation. Moreover, frequency effects, a psychological characteristic of memory, should be observed only from memorized items in the mental lexicon, not from processing of compositional

phrases, which are generated from abstract grammar rules. By contrast, in usage-based approaches, language acquisition results from the interaction between input and human domain-general cognitive processes, and the lexicon and the grammar are not rigidly divided (e.g., Bybee, 2010; Tomasello, 2009). Human linguistic knowledge consists of constructions (e.g., the English transitive) or form–meaning mappings in a particular language (e.g., *subject–verb–object* with the prototypical meaning of one entity acting on the other), ranging from morphemes, words, idioms, and partially lexically filled linguistic patterns (e.g., *The __er the __er*) to fully general and abstract linguistic patterns (e.g., *subject–verb–object*; Goldberg, 2003, 2006). Moreover, constructions with varying sizes and degrees of abstraction are represented and processed based on similar general mechanisms, and speakers can simultaneously retain the representation of both a more general and abstract construction (e.g., *subject–verb–object*) and more lexically specific related forms (e.g., *David kicked the ball, KICKER kicked OBJECT*; Ambridge & Lieven, 2011).

From a usage-based view, L1 construction acquisition results from an accumulation of statistical probabilities and abstraction of regularities out of previous construction encounters. As Ellis (1996, 2002) noted, frequency, a key index of statistical information, promotes entrenchment of words, and the cognitive ability of chunking allows strings of words to be registered in human memory. Subsequent exposure to massive input results in statistical fine-tuning, making the registered relation reflect frequency of encounters. Usage-based theories posit that frequency affects acquisition of constructions at all levels. Thus, L1 speakers should exhibit sensitivity to frequency of not only single words but also compositional phrases. That is, in language comprehension and production, more frequent compositional phrases should be recognized, comprehended, and produced faster (e.g., Ambridge, Kidd, Rowland, & Theakston, 2015; Arnon, 2015; Bybee, 2010).

Despite the similar views regarding the general L1 acquisition mechanism, as Arnon and Snider (2010) pointed out, there are two different assumptions in usage-based approaches regarding compositional phrase representation. One assumption, grounded in work of researchers such as Goldberg (2006) and Wray (2002), is that "phrases that are of sufficient frequency can attain independent representation as a way of making processing more efficient" (Arnon & Snider, 2010, p. 69). There is not yet a clear consensus as to the minimum frequency level (i.e., sufficient frequency), but in this view, there is a qualitative difference between highly frequent phrases (i.e., stored as a whole) and less frequent phrases (i.e., generated or analyzed by the language grammar), and the first type is processed faster than the second. The other assumption, informed by work of researchers such as Bybee and colleagues (Bybee, 2006, 2010; Bybee & Hopper, 2001), is that there is no such qualitative difference. Speakers retain information about previously encountered phrases, and more frequent phrases are more entrenched in speakers' representation. The difference between a more frequent phrase and a less frequent phrase is quantitative (i.e., a difference in the level of activation), resulting from different frequencies of previous phrase encounters. Therefore, phrases need not be highly frequent to be processed faster; relatively more frequent phrases should be processed faster than less frequent phrases regardless of the frequency range. Despite these two different assumptions, compositional phrase frequency effects are accounted for in usage-based approaches.

## Previous work on phrase frequency effects in native English speakers (NSs)

Numerous empirical studies have corroborated the ontological status of compositional phrase frequency effects during receptive language processing in NSs (e.g., Arnon & Snider, 2010; Reali & Christiansen, 2007; Siyanova–Chanturia, Conklin, & van Heuven, 2011; Sosa & Macfarlane, 2002; Tremblay, Derwing, Libben, & Westbury, 2011). Among these, studies by Arnon and Snider (2010) and Hernández, Costa, and Arnon (2016) have provided relatively strong support as indicated by reaction times in a phrasal acceptability task because these two studies used phrases longer than two words (i.e., *don't have to worry*) and controlled for frequencies of subparts of target phrases (e.g., *don't have*). Such a control is important; without it, frequency of the whole target phrases may not have been the only factor determining the results.

Regarding production, studies with NSs have reported phrase frequency effects on phonetic reductions and voice onset time (e.g., Aylett & Turk, 2004; Jannsen & Barber, 2012). Some other studies investigated the role of frequency on production durations of compositional phrases longer than two words in elicited or spontaneous speech. The results have been mixed. First, using an elicited production task, Bannard and Matthew (2008) and Arnon and Priva (2013) reported frequency effects in children and adults, respectively, after frequency of the subparts of the target phrases was controlled for. However, using a similar task, Ellis, Simpson-Vlach, and Maynard (2008) and Tremblay and Tucker (2011) observed no effects. A possible cause for the lack of the effects in these latter two studies may have been methodological. For example, in Ellis et al.'s (2008) study, there was a lack of subpart frequency control, and while some of the stimuli were complete syntactic constituents (e.g., *a great deal of*), others were not (e.g., *and at the*). Regarding spontaneous phrase production durations, there is evidence for compositional phrase frequency effects based on telephone conversations (Arnon & Priva, 2013) and interview speech (Arnon & Priva, 2014) when subpart frequencies were accounted for. Collectively, therefore, research on L1 speech production durations has lent some support to frequency effects on production of compositional phrases longer than two words, but cross-study methodological differences (e.g., type of speech analyzed and substring frequency control) may have been responsible for the inconclusive results.

## Frequency in second language (L2) acquisition

Recent years have seen an increase in usage-based L2 research (e.g., Ellis, 2019; Ellis, Römer, & O'Donnell, 2016; Ortega, 2013; Supasiraprapa, 2018) due to the proposal that, as in L1 acquisition, L2 acquisition is driven by accumulation of statistical probabilities of previously encountered input (e.g., Ellis, 2011, 2013). However, the L2 literature has also documented various L1–L2 acquisition differences, which may lead to different frequency effects between adult NSs and adult non-native English speakers (NNSs). First, the input that L1 learners and adult L2 learners receive may differ in terms of amount and structure (e.g., Littlemore, 2009; Muñoz, 2008). Second, while phrase frequency effects result from implicit learning

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

mechanisms (e.g., Ellis, 2002, 2013) and while L1 and L2 abstract learning patterns may share some similarities (e.g., Wonnacott, 2011), compared to L1 learners, older L2 learners may be less apt at acquiring linguistic patterns implicitly and rely more on explicit learning (e.g., Bley-Vroman, 2009; DeKeyser, 2000). As Wray (2002) observed, L2 learners may also not retain memory about L2 word co-occurrences but instead break phrases into individual words due to various reasons: a lack of necessity to memorize and use L2 word phrases, the typical focus on individual words in adult L2 education, and their mature cognitive development and L1 literacy, which prompt them to break lexical sequences into words. Research on frequency effects on L2 compositional phrase processing is thus interesting because it can potentially shed light on L1–L2 acquisition similarities and differences.

## Previous work on frequency effects in L2 English phrase comprehension and production

To date, empirical research on receptive L2 English compositional phrase processing has lent some support to phrase frequency effects. For example, three studies (Gyllstad & Wolter, 2016; Wolter & Gyllstad, 2013; Wolter & Yamashita, 2018) found that, in phrasal decision tasks, adult NNSs processed English collocations faster when collocation frequency increased. In addition, a study by Sonbul (2015), which used a reading task with concurrent eye-movement registration, reported that relatively proficient adult NNSs demonstrated sensitivity to frequency of English collocations as measured by first pass reading time. In these studies, phrase frequency effects were importantly documented when frequencies of the subparts of the target collocations were controlled for. However, the stimuli consisted of two words (e.g., *middle class* and *pay tax*); as in L1 acquisition, stronger evidence for phrase frequency effects would be from processing of longer phrases. Moreover, in the case of Sonbul's (2015) study, participants had mixed L1 backgrounds; thus, the influence of L1 on L2 collocation processing (Wolter & Gyllstad, 2011, 2013; Wolter & Yamashita, 2018; Yamashita & Jiang, 2010) was not accounted for.

Three other studies investigated compositional three- to five-word sequence processing, offering support (Ellis et al., 2008; Siyanova-Chanturia et al., 2011) or a lack thereof (Valsecchi et al., 2013) based on adult L2 comprehension. These three studies, however, shared the same methodological limitation: a lack of subpart frequency control. Addressing this limitation, Hernández et al. (2016) more recently demonstrated with a phrasal decision task that adult L2 learners' processing of four-word phrases was frequency sensitive, offering relatively strong support for usage-based L2 acquisition.

In terms of frequency effects on L2 English production, Durrant and Schmiit (2010) reported that NNSs' memory retention of collocations, as measured with a productive recall task, increased as a function of the number of times the collocations appeared in a previous training session. In addition, Fernández and Schmitt (2015) reported a significant correlation between NNSs' active recall of English collocations and collocation frequency. While these findings are compatible with a usage-based account, the two studies contained certain methodological limitations. For example, using only an immediate recall task, the former study focused on initial

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

memory retention of L2 word co-occurrences, but not on whether subsequent frequency of collocation encounters (i.e., statistical fine-tuning) leads to frequency effects (Ellis, 1996, 2002). Moreover, in the latter study, frequencies of constituent words of the target collocations were not controlled for, and both studies focused on only two-word L2 collocations.

As for frequency effects on production durations of L2 English phrases longer than two words, relevant research has been scarce. In fact, it seems that there has been only one empirical study, namely, Ellis et al. (2008). Focusing on academic word sequences (e.g., *it should be noted that*), this study observed no frequency effects in both NSs and NNSs in an elicited speech production task. However, as discussed, methodological issues in this study raised a question of whether phrase frequency alone determined participants' production durations. Research on frequency effects on NNSs' compositional phrase production durations thus deserves further investigation.

Perhaps two additional methodological issues in the existing NNS research should be pointed out. First, in studies reporting compositional phrase frequency effects, the typical frequency data source was a large native speaker corpus, including the Corpus of Contemporary American English (Davies, 2013), which was used by researchers such as Wolter and Gyllstad (2013) and Wolter and Yamashita (2018), and the Fisher (Godfrey, Holliman, & McDaniel, 1992) and Switchboard (Cieri, Miller, & Walker, 2004) corpora of American English telephone conversations, which were used by Hernández et al. (2016). While the input that NSs and adult NNSs receive are unlikely to be identical (e.g., Muñoz, 2008), such findings seemed to corroborate the proposal that data from a large and adequately representative corpus should represent the shared regularities of input that all language users have received (Hoey, 2005). Second, most NNS studies reporting phrase frequency effects obtained frequency data from both spoken and written native speaker corpora (e.g., Wolter & Gyllstad, 2013), suggesting that frequency of encounters in written input may affect L2 phrase representation. To date, however, a direct discussion about the role of written texts when compared to the role of speech input in usage-based L2 acquisition has been relatively limited, possibly because L2 acquisition theories originate from child L1 acquisition theories (e.g., Ellis, 2013; Tomasello, 2009). Consequently, at this point, frequency of previous encounters of compositional phrases in spoken L2 input seems most theoretically relevant.

In sum, empirical studies have offered some evidence for L1 and L2 compositional phrase frequency effects, but limitations in the existing studies included a lack of subpart frequency control and scant attention to L2 phrase production. Therefore, addressing these limitations in the following two respective experiments, the present study sought evidence for frequency effects in both receptive and productive processing of L1 and L2 compositional four-word phrases. The current study also constituted the first attempt to investigate such effects in both language comprehension and production in the same adult English L1 and L2 speakers. The research questions were as follows:

1. Are adult NSs and NNSs sensitive to the frequency of compositional four-word phrases in recognition when the frequency of the smaller parts is controlled for?

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

2. Are adult NSs and NNSs sensitive to the frequency of compositional four-word phrases during language production when the frequency of the smaller parts is controlled for?

## Experiment 1

### *Method*

#### *Participants*

Participants were 51 adult NSs (mean age = 20.58, *SD* = 2.93) and 52 adult NNSs (mean age = 23.54, *SD* = 3.93) who were undergraduate or graduate students at a large Midwestern US university. They had no record of hearing impartment or speech difficulty. The NNSs were international students from China and shared the same L1 (Chinese). They can be characterized as being proficient enough to study in an English-speaking environment (minimum internet-based Test of English as a Foreign Language score of 85, mean score = 95.52, *SD* = 6.63). None had an English immersion experience before the age of 10. Moreover, they had a relatively similar length of stay in the United States (range = 2–3 years, mean = 2.61, *SD* = 0.56) and had never lived in another English-speaking country before coming to the United States, except for 1 participant who had stayed in Canada for 5 months.

#### *Materials*

*Target phrases*. This experiment used the phrases from Arnon and Snider (2010) that yielded phrase frequency effects in adult NSs. The target phrases were 28 pairs of phrases, each consisting of two four-word phrases that differed in phrase-frequency (high vs. low) and differed only in the final word (e.g., *don't have to worry* vs. *don't have to wait*). The two variants in each pair were of the same constituent type (e.g., verb phrases or noun phrases). These target pairs were constructed from the Fisher (Godfrey et al., 1992) and Switchboard (Cieri et al., 2004) corpora of American English telephone conversations, which together contained around 20 million words. The phrases are therefore common in general English conversations. At the end of Experiment 2 in the current study, when asked to identify any component words they did not know, participants who were NNSs indicated that they knew all the words in the target phrases.

As in Arnon and Snider's (2010) study, the 28 target pairs were from a high- and a low-cutoff bin, which consisted of 16 and 12 target pairs, respectively. In the high cutoff bin, based on the frequencies derived from Fisher and Switchboard, each high-frequency variant occurred at least 12.00 times per million words, while each low-frequency variant occurred less often. By contrast, in the low cutoff bin, in each pair, the high-frequency variant occurred at least 1.00 time per million words, while the low-frequency variant occurred less frequently. The purpose of incorporating the two cutoff bins was twofold. First, Arnon and Snider (2010) observed frequency effects from adult NSs in a phrasal acceptability task in both bins. In light of the two previously discussed proposals about usage-based compositional phrase representation, Arnon and Snider (2010) argued that phrases do not

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

**Table 1.** Examples of the 28 target pairs

| Cutoff bin | Phrases | Frequency condition | Frequency (per million words) |
|---|---|---|---|
| High | don't have to worry | high | 20.35 |
| | don't have to wait | low | 2.00 |
| | I don't know why | high | 47.85 |
| | I don't know who | low | 11.60 |
| Low | a lot of rain | high | 6.00 |
| | a lot of blood | low | 0.25 |
| | don't have any money | high | 2.80 |
| | don't have any place | low | 0.45 |

need to be highly frequent (i.e., have frequency above the cutoff point in the high cutoff bin) to be processed faster. Therefore, there was no direct evidence that highly frequent phrases were qualitatively different from phrases with lower frequencies. Instead, because frequency effects were similarly observed in both cutoff bins, the researchers argued that the differences in reaction times to the target phrases should have resulted from relative quantitative differences (i.e., different frequencies of previous phrase encounters), regardless of the frequency range (Bybee, 2006, 2010; Bybee & Hopper, 2001). The current experiment therefore investigated whether Arnon and Snider's (2010) results are replicable in adult NSs. The second reason was specifically related to the NNSs, who presumably had had less exposure to English input. These learners may exhibit frequency effects only with target phrases in the high cutoff bin because they may have been exposed to many instances of highly frequent phrases. In contrast, the NNSs may not have had much exposure to the target phrases in the low cutoff bin and thus may not demonstrate frequency effects in this bin. In short, an additional goal of having the two cutoff bins was to illuminate whether adult NNSs had stored sufficient accumulated statistics information (i.e., frequency of occurrences) through previous L2 exposure to demonstrate frequency effects in both cutoff bins.

Table 1 shows examples of the target phrases. The first two pairs are examples from the high cutoff bin. The frequency cutoff point for classifying a phrase as a high- or low-frequency phrase in this bin (12.00 times per million words) was slightly higher than that in Arnon and Snider's (2010) study (10.00 times per million words) because in the current study the frequencies of the target phrases, although obtained from the same corpora, were found to be slightly higher than the frequencies reported in the original study. The cutoff point was therefore slightly increased so that all the low-frequency phrases in the original study were still considered low-frequency phrases in the current study (e.g., in the current study *I don't know who*, a low-frequency phrase, was found to occur 11.60 times per million words). The third and fourth pairs are examples from the low cutoff bin, in which the cutoff point for classifying a phrase as a high- or a low-frequency phrase (1.00 time per million words) is the same as that in Arnon and Snider's (2010) study. Thus, the

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

classification of a phrase as having high or low frequency was meaningful within each cutoff bin, not across all the 28 stimuli pairs. The 28 target pairs and their frequencies are listed in Appendix A.

In the high cutoff bin and based on the frequency data derived from Fisher and Switchboard, the mean frequencies of the high- and low-frequency variants across the 16 target pairs were 25 occurrences (Min = 12.00, Max = 53.15, *SD* = 12.97) and 4.87 occurrences (Min = 0.70, Max = 11.60, *SD* = 3.93) per million words, respectively. Across these pairs, frequencies between high- and low-frequency phrases differed significantly, $t$ (30) = –10.76, $p < .001$. In the low cutoff bin, the mean frequencies of the high- and low-frequency variants across the 12 target pairs were 4.68 occurrences (Min = 1.85, Max = 12.60, *SD* = 3.18) and 0.27 occurrences (Min = 0.05, Max = 0.55, *SD* = 0.14) per million words, respectively. Across these 12 pairs, frequencies between high- and low-frequency phrases also differed significantly, $t$ (22) = –4.81, $p < .001$. Therefore, in each cutoff bin, high-frequency phrases occurred significantly more often than low-frequency phrases.

Regarding the subparts of the phrases, each pair (e.g., *don't have to worry* vs. *don't have to wait*) differed only in three subparts: the last word (e.g., *worry* vs. *wait*), the last two words (e.g., *to worry* vs. *to wait*), and the last three words (e.g., *have to worry* vs. *have to wait*). To observe effects of the four-word sequence more clearly, the frequencies of these subparts will later be entered as control variables in the analysis. Finally, the high- and low-frequency phrases in each cutoff bin did not differ in terms of the plausibility of meaning. This was attested by results from a plausibility rating task completed by a group of NSs in Arnon and Snider's (2010) study. This is not surprising as all the phrases are possible and meaningful in English.

*Fillers*. Besides the 56 target phrases, there were 80 four-word-phrase fillers of two types. As in Arnon and Snider's (2010) study, the first were 12 possible phrases (e.g., *buy a new dress*), and the second were 68 impossible phrases. Among the latter, 75% had a wrong word order (e.g., *she a has boat*), while 25% were impossible due to an incorrect preposition use (e.g., *afraid to the dark*). An attempt was made to avoid an overlap between words in the target phrases and the fillers. In total, there was an equal number of possible phrases (54 target phrases plus 12 fillers) and impossible phrases (68 fillers).[1]

### *Procedure*

The stimuli were divided into two blocks: A and B. One variant from each of the 28 target pairs was randomly assigned to only one block to minimize a possible repetition effect resulting from participants seeing the identical first three words in the two variants of the same target pair in the same block. Each block thus consisted of 14 high-frequency variants and 14 low-frequency variants from the target pairs. Fillers were randomly and equally assigned to each block such that, in total, each block contained 6 fillers, which are possible English word sequences, and 34 fillers, which are impossible sequences. The total number of phrases in each block was thus 68, with half of the phrases being grammatical and the other half ungrammatical. The stimuli were presented in a random order.

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

The experiment was run on Superlab (Cedrus Corporation, 2006). Only informed that the current study investigates English phrase comprehension and production, each participant sat in a quiet room in front of a computer screen and completed a phrasal decision task, in which they saw four-word phrases in the center of the screen, one at a time, and were asked to press a *YES* button or a *NO* button, which were equally positioned on a keypad, to judge whether the phrases were possible English word sequences. Two different instructions and keypads were used with right-handed and left-handed participants. The participants were instructed to make their judgment as fast as they could while still being accurate. Participants also completed a short practice section, in which they saw examples of both possible and impossible sequences, before the actual experiment began. During the experiment, participants first saw a plus sign in the center of the screen for eye fixation. The sign lasted 333 ms and was followed by a long blank screen presented for 50 ms. A phrase then appeared and remained on the screen until a button press. The phrases appeared one at a time and in their entirety (font: Arial; size: 36; position: center). Words were in the lowercase except the first-person personal pronoun and proper names.

This experiment followed a within-subject counterbalanced design. Half of the participants in each group were randomly assigned to complete Block A first, while the other half completed Block B first. Each participant completed both blocks and thus saw the two variants from each target pair across the two blocks. Between the two blocks, there was a break during which participants completed the first part of their background questionnaire. The break was included to further reduce possible repetition effects. The whole experiment took approximately 20 minutes.

### *Analysis*

Because the frequency cutoff point in each cutoff bin differed, a separate analysis for each cutoff bin was conducted. Only reaction times for the target phrases were analyzed. One NS and 1 NNS were excluded due to relatively low levels of judgment accuracy of 88% and 77%, respectively; the minimum accuracy levels to exclude participants in the two respective groups were 93% and 89%. Excluded as well were 2 NNSs who occasionally stopped during the experiment to do unrelated activities. The remaining 50 NSs and 49 NNSs had a mean accuracy rate of 98% ($SD = 0.02$) and 97% ($SD = 0.03$), respectively. Therefore, both groups did not seem to have any difficulty doing the task. Incorrect responses were also excluded. Moreover, in each group, reaction times exceeding +/–2 *SD* from the group mean in each frequency condition in each cutoff bin were removed, resulting in a removal of 3% and 4% of the correct responses from the NSs and the NNSs, respectively.

The data were analyzed with mixed-effects regression models (Baayen, Davidson, & Bates, 2008; Bates, 2010). Models were run in R (R Core Team, 2015) with the statistics package *lme4* (Bates, 2010; Bates, Mächler, Bolker, & Walker, 2015). Reaction times were log transformed to reduce skewness of the data.[2] Fixed effects included phrase frequency condition (high/low), participant group (NS/NNS), the interaction between these two variables, and control variables: block order (i.e., whether the participant completed that block as the first or the second block), the number of characters of the target phrases, and the frequencies of the subparts

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

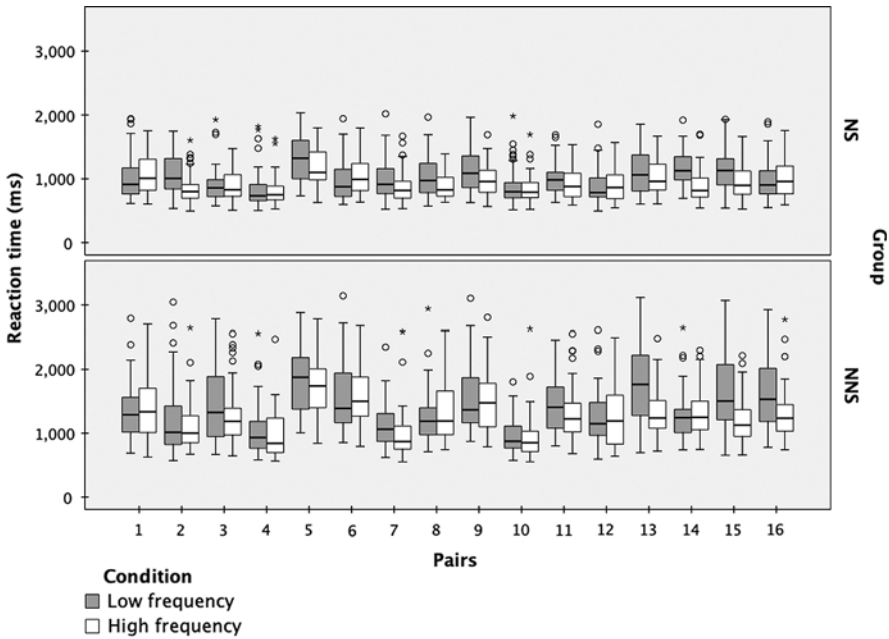


**Figure 1.** Boxplots for reaction times (ms) for each stimuli pair in the high cutoff bin. The order of the pairs on the *x*-axis is based on the frequency of the high-frequency member of each pair; the high-frequency member of the pair on the far left has the lowest frequency among all the high-frequency members of all pairs. NS = native English speaker. NNS = non-native English speaker.

that differ in each pair (e.g., frequencies of *to worry* and *to wait* in the pair *don't have to worry* and *don't have to wait*). The interaction between block order and phrase frequency was also included. Continuous predictors were standardized to illustrate the magnitude of such predictors. Moreover, participant and item were included in the model as random intercepts. Wald statistics (Type II) were used to determine the significance of the effects (Fox, 2008; Fox & Weisberg, 2011).

In the analysis, model comparisons suggested that adding different reaction time slopes for participant and item random effects did not significantly improve the models, either for the high cutoff bin ($p = .179$) or the low cutoff bin ($p = .185$). Consequently, the final model for each cutoff bin allowed for a random intercept for each participant and each test item, but a by-participant random slope for phrase frequency was not included. In each final model, the variance inflation factor (VIF) scores for the continuous predictors were well below the threshold level of 10 (Loewen & Plonsky, 2016; Myers, 1990), and visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality.

### *Results*

Figure 1 and Figure 2 show boxplots of reaction times from the phrasal acceptability judgment task in milliseconds for each member of the pairs in the high and low cutoff bins, respectively. Generally, participants in both groups were faster when

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

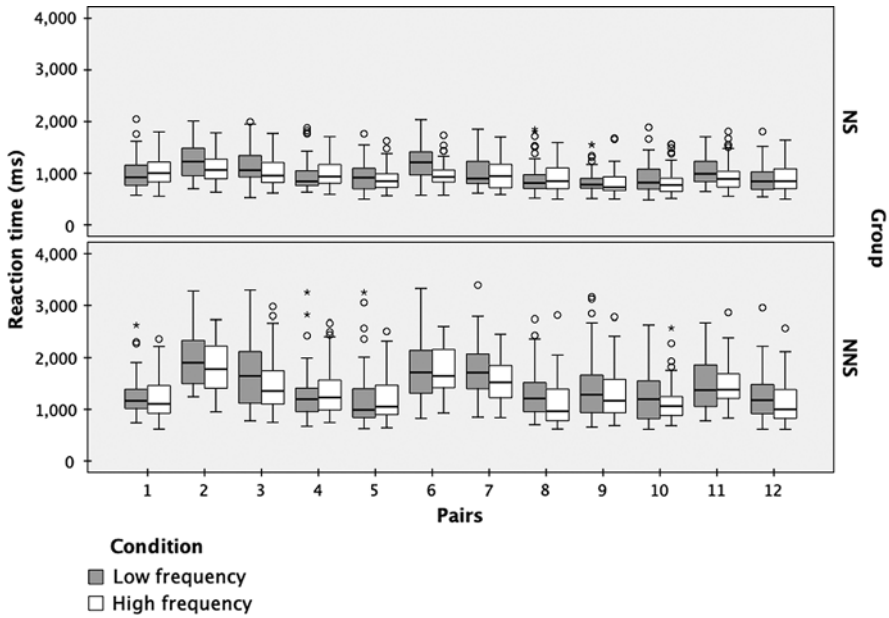


**Figure 2.** Boxplots for reaction times (ms) for each stimuli pair in the low cutoff bin. The order of the pairs on the *x*-axis is based on the frequency of the high-frequency member of each pair; the high-frequency member of the pair on the far left has the lowest frequency among all the high-frequency members of all pairs. NS = native English speaker. NNS = non-native English speaker.

judging the high-frequency phrase in each pair, and the NSs were faster than the NNSs.

Table 2 shows the average reaction times from the two participant groups. The last column indicates the difference in reaction time between the high-frequency condition and the low-frequency condition (the baseline frequency category) in each cutoff bin in each participant group. In the high cutoff bin, in which the cutoff point for classifying a phrase as a high-frequency phrase or a low-frequency phrase was 12.00 occurrences per million, the native speaker mean reaction time for high-frequency phrases was about 84 ms shorter than the mean reaction time for low-frequency phrases. This difference was greater than the difference of 60 ms in Arnon and Snider's (2010) study, in which native speaker mean reaction time for high- and low-frequency phrases were 1040 ms and 1100 ms, respectively. In the low cutoff bin, in which the cutoff point was 1.00 occurrence per million, the mean reaction time for high-frequency phrases was approximately 55 ms shorter than the mean reaction time for low-frequency phrases in the native speaker group. This was smaller than the 66-ms difference in Arnon and Snider's (2010) study, in which respective mean reaction time for high- and low-frequency phrases were 1040 ms and 1100 ms. Moreover, while the NNSs were generally slower than the NSs, these NNSs also reacted faster to high-frequency phrases than to low-frequency phrases. In the high and the low cutoff bins, respectively, the mean reaction times for high-frequency phrases were about 116 ms and 94 ms shorter than the mean reaction times for low-frequency phrases.

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

**Table 2.** Average reaction times in milliseconds from the phrasal acceptability judgment task (*SD* in parentheses)

| | Phrase frequency | | Frequency effects |
|---|---|---|---|
| | Low | High | (High – Low) |
| NS ($N = 50$) | | | |
| High cutoff bin | 1,029.06 (339.52) | 944.94 (279.23) | –84.12 |
| Low cutoff bin | 1,003.70 (328.08) | 958.40 (288.24) | –55.30 |
| NNS ($N = 49$) | | | |
| High cutoff bin | 1,399.37 (556.35) | 1,283.86 (476.90) | –115.51 |
| Low cutoff bin | 1,462.25 (605.05) | 1,368.19 (504.81) | –94.06 |

*Note*: NS = native English speaker. NNS = non-native English speaker.

**Table 3.** Mixed-effects model results for the high cutoff bin in the phrasal acceptability task

| Predictors | β | 95% CI | $10^{\beta}$ | *SE* | *p* |
|---|---|---|---|---|---|
| Phrase frequency (baseline = low) | –0.04 | [–0.07, –0.01] | 0.91 | 0.02 | .008 |
| Group (baseline = NS) | 0.12 | [0.09, 0.16] | 1.33 | 0.02 | <.001 |
| Block order (baseline = first) | –0.04 | [–0.05, –0.03] | 0.91 | 0.004 | <.001 |
| Phrase frequency × Group | –0.002 | [–0.01, 0.02] | 1.00 | 0.01 | .842 |
| Phrase frequency × Block order | –0.01 | [–0.02, 0.01] | 0.99 | 0.01 | .453 |
| Number of characters | 0.04 | [0.02, 0.06] | 1.10 | 0.01 | <.001 |
| Log frequency of last word | –0.002 | [–0.03, 0.02] | 1.00 | 0.01 | .842 |
| Log frequency of last two words | –0.01 | [–0.03, 0.02] | 0.98 | 0.01 | .444 |
| Log frequency of last three words | 0.01 | [–0.01, 0.03] | 1.02 | 0.01 | .269 |

*Note*: $R^2$ marginal = .26. $R^2$ conditional = .54. *SE* = standard error. NS = native English speaker.

Table 3 reports results from the regressions for the high cutoff bin. There were significant main effects of phrase frequency, $\chi^2$ (1) = 7.13, $p = .008$, participant group, $\chi^2$ (1) = 61.65, $p < .001$, block order, $\chi^2$ (1) = 115.45, $p < .001$, and the number of characters in the target phrases, $\chi^2$ (1) = 21.74, $p < .001$. The regression coefficient (β) for each predictor indicates the change in reaction times on the base-10 logarithmic scale as a result of a 1-unit change in the predictor. In the case of binary predictors, the exponential value ($10^{\beta}$) expresses the average multiplicative change in reaction time between the nonreference category (e.g., high) and the reference category (e.g., low). Similarly, for continuous predictors, which were standardized, the exponential value expresses the average multiplicative change in reaction time associated with a 1 *SD* change in the predictor.

As Table 3 shows, both groups demonstrated sensitivity to phrase frequency. On average, reaction times to high-frequency phrases were 0.91 times the reaction times to low-frequency phrases. That is, participants were about 9% faster when judging

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

**Table 4.** Mixed-effects model results for the low cutoff bin in the phrasal acceptability task

| Fixed effects | β | 95% CI | $10^{\beta}$ | *SE* | *p* |
|---|---|---|---|---|---|
| Phrase frequency (baseline = low) | −0.04 | [−0.10, −0.004] | 0.90 | 0.02 | .029 |
| Group (baseline = NS) | 0.15 | [0.13, 0.19] | 1.42 | 0.02 | <.001 |
| Block order (baseline = first) | −0.06 | [−0.07, −0.04] | 0.88 | 0.01 | <.001 |
| Phrase frequency × Group | −0.001 | [−0.02, 0.01] | 0.99 | 0.01 | .696 |
| Phrase frequency × Block order | 0.01 | [−0.01, 0.02] | 1.01 | 0.01 | .531 |
| Number of characters | 0.04 | [0.02, 0.05] | 1.09 | 0.01 | <.001 |
| Log frequency of last word | 0.02 | [−0.003, 0.04] | 1.06 | 0.01 | .053 |
| Log frequency of last two words | −0.02 | [−0.05, 0.01] | 0.96 | 0.02 | .250 |
| Log frequency of last three words | 0.01 | [−0.01, 0.04] | 1.02 | 0.01 | .442 |

*Note*: $R^2$ marginal = .31. $R^2$ conditional = .56. *SE* = standard error. NS = native English speaker.

the acceptability of high-frequency phrases. Moreover, the NNSs were on average 33% slower than the NSs, and participants were on average 9% faster when they did the second experiment block, whether it was Block A or Block B, possibly due to greater task familiarity. While the counterbalanced design equally distributed the familiarity effects between the two blocks, with block order as a control variable, the portion of variance in reaction time due to task familiarity was accounted for statistically.

In addition, a 1 *SD* increase in the number of characters in the target phrases corresponded to about a 10% increase in reaction times, indicating that the participants had to spend more time reading the phrases. The interaction between phrase frequency and group, $\chi^2$ (1) = 0.04, $p$ = .842, and the interaction between phrase frequency and block order, $\chi^2$ (1) = 0.56, $p$ = .453, were not significant, and neither were the frequencies of the last word, $\chi^2$ (1) = 0.04, $p$ = .842, the last two words, $\chi^2$ (1) = 0.59, $p$ = .444, and the last three words, $\chi^2$ (1) = 1.22, $p$ = .269. As in previous similar research (Gyllstad & Wolter, 2016), the MUMIn function in R was used to obtain $R^2$ values. This function provides two $R^2$ values: marginal and conditional. The former is associated with the fixed effects, listed in this table, and the latter reflects the fixed and the random effects combined. In this model, the fixed effects and the random effects together explain about 54% of the variance in the participants' reaction times.

As Table 4 shows, a similar result pattern was obtained from the low cutoff bin. The effects of phrase frequency, $\chi^2$ (1) = 4.78, $p$ = .029, participant group, $\chi^2$ (1) = 85.79, $p$ < .001, block order, $\chi^2$ (1) = 125.92, $p$ < .001, and the number of characters in the target phrases, $\chi^2$ (1) = 19.38, $p$ < .001, were significant. Therefore, both groups were sensitive to phrase frequency. On average, they were 10% faster when judging high-frequency phrases, and the NNSs were approximately 42% slower than the NSs. Reaction times in the second experiment block were on average 12% shorter than in the first block, and a 1 *SD* increase in the number of characters in the target phrases corresponded to an approximately 9% increase in reaction times.

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

The interaction between phrase frequency and group, $\chi^2$ (1) = 0.15, $p$ = .696, and the interaction between phrase frequency and block order, $\chi^2$ (1) = 0.39, $p$ = .531, were not significant. Moreover, frequencies of the last two words, $\chi^2$ (1) = 1.32, $p$ = .250, and the last three words, $\chi^2$ (1) = 0.59, $p$ = .442, did not reach significance, while frequencies of the last words, $\chi^2$ (1) = 3.75, $p$ = .053, was close to being significant.

### *Discussion*

Experiment 1 tested the prediction in usage-based approaches that adult NSs and NNSs should demonstrate frequency effects during receptive compositional phrase processing (e.g., Bybee, 2010; Ellis, 2011). Such effects were previously observed from adult NSs based on compositional phrases consisting of more than two words in the study by Arnon and Snider (2010) and a recent study by Hernández et al. (2016). This latter study was also the first that reported frequency effects based on such phrases in adult NNSs. The results from the current experiment were in line with those from the two preceding studies; in both stimuli cutoff bins and both participant groups, reaction time for high-frequency phrases was significantly shorter than reaction time for low-frequency phrases. Because subpart frequencies, block order, and the number of characters in the target phrases were controlled for, the frequency effects should have resulted from higher whole phrase frequency.

The findings thus lend support to the proposal in usage-based approaches that L1 acquisition is based on an accumulation of statistical information in previously encountered input and that multiword processing should be frequency sensitive (e.g., Arnon, 2015; Bybee, 2010; Ellis, 2011). Regarding the two cutoff bins, the findings are in line with those from a previous study by Arnon and Snider (2010). In that study, NSs had significantly shorter reaction time to higher frequency phrases in both the high and the low cutoff bins, leading the researchers to argue that NSs do not process only highly frequent phrases (i.e., those with frequencies above the cutoff point in the high cut off bin) faster. That is, their findings do not seem to support the proposal that there is a high-frequency threshold and that only compositional phrases with a frequency above this threshold level are stored as a whole and are processed faster than less frequent phrases, which are not stored as a whole but are analyzed or computed based on language grammar (e.g., Goldberg, 2003, 2006; Wray, 2002). Rather, according to Arnon and Snider (2010), the differences in reaction times to higher and lower frequency phrases in both cutoff bins should support the proposal that more frequent phrases are more entrenched in speakers' representation, regardless of the frequency range. Therefore, a difference between processing a more frequent phrase and a less frequent phrase is quantitative, resulting from relative differences in frequencies of previous encounters (i.e., different levels of activation). The current experiment, therefore, replicated Arnon and Snider's (2010) findings and extended these findings to adult NNSs.

As for NNSs, the results in the current experiment suggested that the amount of exposure to compositional multiword phrases that the NNSs had accumulated was sufficient for them to exhibit sensitivity to phrase frequencies derived from large NS spoken corpora, whether the phrases are in the low or high end of the frequency range (i.e., high or low cutoff bin). This is in line with the findings in

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

Hernández et al.'s (2016) study and appears to corroborate usage-based researchers' proposal that L2 acquisition may also be based on the general mechanism operating in L1 acquisition (e.g., Ellis, 2011, 2013). That is, adult NNSs can also retain memory about word co-occurrences, and the speed of receptive phrase processing increases as a function of frequency of previous phrase encounters. This does not mean that the NSs and the NNSs in the current experiment had received identical English input. Instead, similar to previous English L2 studies using NS corpora (e.g., Gyllstad & Wolter, 2016; Wolter & Gyllstad, 2013), the results from the current experiment support the argument that a sufficiently large NS corpus should represent the common regularities of input all speakers have been exposed to (Hoey, 2005).

Moreover, like Hernández et al.'s (2016) study, the current experiment did not find a significant interaction between participant group and phrase frequency. This differs from the consistent finding in single-word recognition research (e.g., Diependaele, Lemhöfer, & Brysbaert, 2013; Whitford & Titone, 2012), according to which frequency effects were observed from NSs and English L2 speakers but were stronger in the latter group, as indicated by such an interaction. The interaction has led to a proposal that, due to lower English proficiency and less English input, English words are less well entrenched in English L2 speakers' mental representation than in NSs' representation. Therefore, processing of English L2 words generally requires more effort, but particularly greater effort is required when L2 words have low frequency. Consequently, the difference in processing high- and low-frequency English words is more pronounced in L2 speakers than in L1 speakers (e.g., Diependaele et al., 2013). In Hernández et al.'s (2016) study, besides a phrasal acceptability task, the researchers conducted a lexical decision task in which the same NSs and NNSs in the phrasal acceptability task judged whether strings of letters on a computer screen were English words. The lexical decision task revealed frequency effects in both groups, but stronger effects in the learner group, as indicated by the interaction between group and word frequency. Hernández et al. (2006) speculated that, in the phrasal acceptability task, stronger frequency effects in the NNSs may also exist but were not observed because the mean frequency differences between high- and low-frequency phrases were too low compared to the mean frequency differences between high- and low-frequency words in the lexical decision task. In the current experiment, as in Hernández et al.'s (2006) study, a part of the stimuli from Arnon and Snider (2010) was used and no interaction between participant group and frequency was observed. Future studies can therefore investigate the interaction between participant group and frequency using phrases with a wider frequency range.

## Experiment 2

This experiment tested frequency effects on production of four-word compositional phrases. An elicited oral production task was used because this task demonstrated phrase frequency effects in two previous studies (Arnon & Priva, 2013; Bannard & Matthew, 2008) and helped increase the comparability between the two experiments in the current study (i.e., elicitation of the same target phrases). Moreover, as in these previous studies, language production was operationalized as the phonetic

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

durations of the first three words in the target phrases, with the assumption that the same words should be produced faster in a more frequent phrase (e.g., *don't have to worry*) than in a less frequent phrase (e.g., *don't have to wait*).

## Method

### Participants

To control for individual differences across the two experiments, the same participants in Experiment 1 completed the present experiment. Moreover, to minimize participants' familiarity with the target phrases, the participants completed this experiment at least 2 days after Experiment 1. All participants completed Experiment 1 before Experiment 2 so that their exposure to the target phrases in Experiment 2 did not influence their acceptability judgments of these phrases in Experiment 1.

### Materials

The same 28 target pairs from Experiment 1 were divided into two blocks: C and D. As in Experiment 1, one variant from each target pair was randomly assigned to each block, and each block consisted of 14 high-frequency variants and 14 low-frequency variants from the target pairs. Each block therefore contained 28 target phrases. In addition, in each block, there were 28 fillers, which were possible English phrases. Thus, each block contained 56 phrases. The fillers in this experiment consisted of grammatical fillers from Experiment 1 (e.g., *buy a new dress*) and grammatical counterparts (e.g., *afraid of the dark*) of ungrammatical fillers from Experiment 1 (e.g., *afraid to the dark*). The purpose was to ensure the comparability of the fillers in the two experiments so that the target items, which were the same in both experiments, would not stand out. Lexical overlap between fillers and the target phrases was also minimized.

### Procedure

The experiment was run on PsychoPy (Peirce, 2007). Each participant sat in a quiet room and completed a phrase elicitation task in front of a computer screen. As in a similar previous study by Arnon and Priva (2013), which demonstrated phrase frequency effects in NSs, to investigate the effects of the frequency of a whole target phrase, the instruction told participants to read the phrase as soon as the phrase disappeared (i.e., after seeing the whole phrase). In addition, as in previous similar research (e.g., Ellis et al., 2008; Janssen & Barber, 2012), participants were instructed to read the phrase as fast and as accurately as they could. Each target phrase appeared on the screen one at a time and in its entirety (font: Arial; size: 36; position: center) for a fixed amount of time (1700 ms). Words were in the lowercase, except for the first-person personal pronoun and proper nouns. Phrases were presented in a random order. Based on a pilot study, the interval between the end of a phrase presentation and the time the next phrase appeared was set at 2500 ms because this duration was found to be sufficiently long for speech production of both participant groups. There were also six practice items at the beginning of each experiment block, and participants were informed that their voice would be recorded. PsychoPy started recording participants' production from the moment each phrase

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

disappeared from the computer screen to the moment the following phrase appeared. Thus, the recorded duration for each phrase was 2500 ms.

A within-subject counterbalanced design was again used to control for participants' individual variability in processing phrases in the two blocks. Participants in each group were randomly and equally divided to complete either Block C or Block D first. Each participant completed both blocks, which were separated by a break in which they filled out the second part of the background questionnaire. The whole experiment took approximately 25 minutes. Participants received 20 USD after the completion of this experiment.

### *Analysis*

PRAAT (Boersma & Weenink, 2010) and Dartmouth Linguistic Automation (DARLA) web interface (Reddy & Stanford, 2015) were used to identify the production duration of the target segment. The identification was carried out twice to maximize accuracy. One NS and 1 NNS did not complete this experiment as they did not return to the lab. Moreover, another NNS, who did not follow the directions (i.e., deliberately emphasized words), was excluded. Another NNS was removed as an outlier because the participant's mean production duration was longer than +2 *SD* from the NNSs' group mean in several conditions. The remaining participants consisted of 50 NSs and 49 NNSs. The mean production accuracy in the two respective groups was 99% ($SD = 0.01$) and 97% ($SD = 0.03$). Incorrect and incomplete responses were also removed. Finally, in each group, production durations outside +/–2 *SD* from the group mean in each frequency condition in each cutoff bin were removed, resulting in an exclusion of about 4% of the correct responses from each group.

Regression models similar to those in Experiment 1 were run, with block order, the number of syllables in the target segment, and subpart frequencies as control variables. Only in the model for the low cutoff bin, log frequencies of last word and the last two words of the target phrases had high VIF scores of 10.81 and 11.22, respectively. The two VIF values were very similar, and removing either of the effects that yielded the high VIF values reduced the VIFs of the remaining continuous predictors to below 10. In the regression model for the high cutoff bin, which will be described in the following Results section, log frequency of the last two words was a significant predictor of participants' production durations. Therefore, in the model for the low cutoff bin, log frequency of the last two words was retained and log frequency of the last word was removed from the analysis. After the removal, the VIF values for all the remaining continuous variables were below 10. Moreover, a by-participant random slope for frequency did not significantly improve the models, either in the high cutoff bin ($p = .809$) or the low cutoff bin ($p = .963$). Thus, the random slope was not included. Finally, in each final model, residual plots for the regression model suggested no obvious deviations from linearity, homoscedasticity, or normality.

## Results

Figure 3 and Figure 4 show boxplots of production durations of the target segments in milliseconds for each target pair in the high and the low cutoff bin, respectively.

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

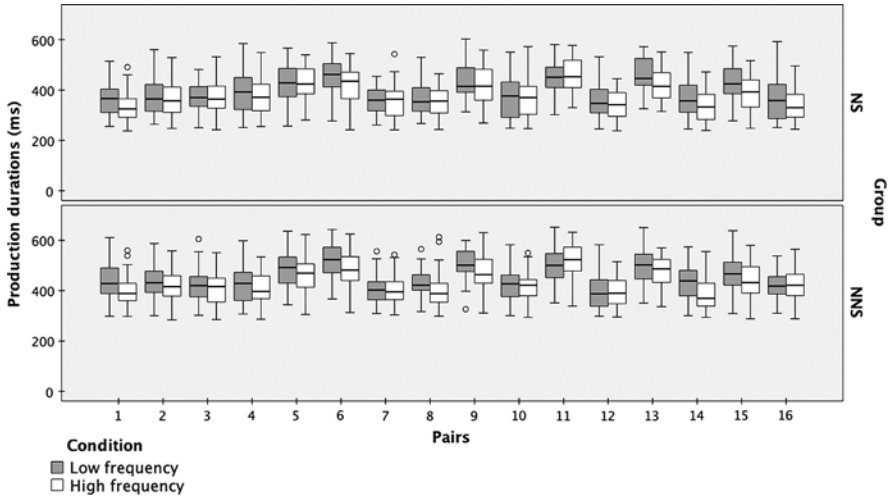


**Figure 3.** Boxplots for production durations (ms) of the target segments for each stimuli pair in the high cutoff bin. The order of the pairs on the *x*-axis is based on the frequency of the high-frequency member of each pair; the high-frequency member of the pair on the far left has the lowest frequency among all the high-frequency members of all pairs. NS = native English speaker. NNS = non-native English speaker.

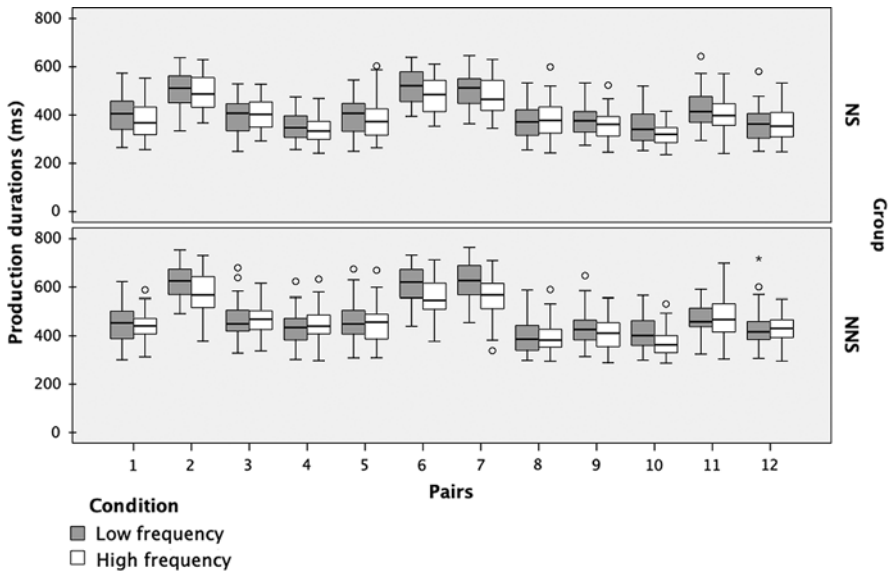


**Figure 4.** Boxplots for production durations (ms) of the target segments for each stimuli pair in the low cutoff bin. The order of the pairs on the *x*-axis is based on the frequency of the high-frequency member of each pair; the high-frequency member of the pair on the far left has the lowest frequency among all the high-frequency members of all pairs. NS = native English speaker. NNS = non-native English speaker.

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

**Table 5.** Average production durations of the target segments in milliseconds from the elicited production task (SD in parentheses)

| | Phrase frequency | | Frequency effects |
|---|---|---|---|
| | Low | High | (High – Low) |
| NS ($N = 50$) | | | |
| High cutoff bin | 397.23 (79.05) | 379.15 (75.74) | –18.09 |
| Low cutoff bin | 416.32 (90.53) | 402.75 (88.71) | –13.57 |
| NNS ($N = 49$) | | | |
| High cutoff bin | 450.23 (75.99) | 432.23 (74.23) | –18.00 |
| Low cutoff bin | 483.67 (107.91) | 466.96 (93.44) | –16.71 |

*Note*: NS = native English speaker. NNS = non-native English speaker.

Overall, the target segment in the high-frequency member of each pair was produced faster than the target segment in the low-frequency member, and the NSs were faster than the NNSs.

Table 5 shows the average production duration for the target segments in each condition. In the NS group, in both high and low cutoff bins, production durations for the target segments inside high-frequency phrases were shorter than production durations for the target segments inside low-frequency phrases. That is, in the two respective bins, the mean production durations for the target segments in high-frequency phrases were about 18 ms and 14 ms shorter than the mean production durations for the target segments in low-frequency phrases. Similarly, in the NNS group, the mean production durations for the target segments in high-frequency phrases were approximately 18 ms and 17 ms shorter than the mean production durations for target segments in low-frequency phrases, respectively.

However, as shown in Table 6, results from the regression models for the high cutoff bin suggested that phrase frequency did not significantly predict production durations of the target segments. Although the coefficient for phrase frequency was negative as expected, phrase frequency effects did not reach significance, $\chi^2$ (1) = 2.35, $p = .125$. There were, however, significant main effects of participant group, $\chi^2$ (1) = 27.93, $p < .001$, block order, $\chi^2$ (1) = 16.76, $p < .001$, and the number of syllables in the target segments, $\chi^2$ (1) = 12.77, $p = < .001$. That is, the NNSs were on average 14% slower than the NSs. Participants were on average 2% faster when they did the second experiment block possibly due to greater task familiarity. In addition, a 1 *SD* increase in the number of syllables in the target segments led to about a 5% increase in production durations. The interaction between phrase frequency and group, $\chi^2$ (1) = 0.47, $p = .493$, and the interaction between phrase frequency and block order, $\chi^2$ (1) = 0.62, $p = .432$, were not significant. The frequencies of the last word, $\chi^2$ (1) = 0.28, $p = .595$, and the last three words, $\chi^2$ (1) = 1.22, $p = .270$, also did not reach significance. However, the effects of frequency of the last two words in the target segment were significant, $\chi^2$ (1) = 6.60, $p = .010$. The negative coefficient for this predictor suggested that a 1 *SD* increase in the log frequency of the last two words in the target phrases (e.g., *to worry* in *don't have to*

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

**Table 6.** Mixed-effects model results for the high cutoff bin in the elicited production task

| Fixed effects | β | 95% CI | $10^{\beta}$ | *SE* | *p* |
|---|---|---|---|---|---|
| Phrase frequency (baseline = low) | −0.02 | [−0.05, 0.01] | 0.95 | 0.01 | .125 |
| Group (baseline = NS) | 0.06 | [0.03, 0.08] | 1.14 | 0.01 | <.001 |
| Block order (baseline = first) | −0.01 | [−0.015, −0.004] | 0.98 | 0.003 | <.001 |
| Phrase frequency × Group | 0.002 | [−0.004, 0.01] | 1.01 | 0.004 | .493 |
| Phrase frequency × Block order | 0.003 | [−0.004, 0.01] | 1.01 | 0.004 | .432 |
| Number of syllables in target segment | 0.02 | [0.01, 0.03] | 1.05 | 0.01 | <.001 |
| Log frequency of last word | 0.004 | [−0.01, 0.02] | 1.01 | 0.01 | .595 |
| Log frequency of last two words | −0.02 | [−0.04, −0.01] | 0.95 | 0.01 | .010 |
| Log frequency of last three words | 0.01 | [−0.01, 0.02] | 1.02 | 0.01 | .270 |

*Note*: $R^2$ marginal = .22. $R^2$ conditional = .70. *SE* = standard error. NS = native English speaker.

**Table 7.** Mixed-effects model results for the low cutoff bin in the elicited production task

| Fixed effects | β | 95% CI | $10^{\beta}$ | *SE* | *p* |
|---|---|---|---|---|---|
| Phrase frequency (baseline = low) | −0.03 | [−0.06, −0.004] | 0.94 | 0.02 | .073 |
| Group (baseline = NS) | 0.07 | [0.04, 0.09] | 1.17 | 0.01 | <.001 |
| Block order (baseline = first) | −0.003 | [−0.01, −0.004] | 0.99 | 0.003 | <.001 |
| Phrase frequency × Group | 0.002 | [−0.004, 0.01] | 1.01 | 0.004 | .527 |
| Phrase frequency × Block order | −0.01 | [−0.015, −0.001] | 0.98 | 0.004 | .046 |
| Number of syllables in target segment | 0.06 | [0.05, 0.07] | 1.15 | 0.01 | <.001 |
| Log frequency of last two words | −0.004 | [−0.02, 0.01] | 0.99 | 0.01 | .636 |
| Log frequency of last three words | 0.01 | [−0.01, 0.03] | 1.02 | 0.01 | .325 |

*Note*: $R^2$ marginal = .45. $R^2$ conditional = .81. Log frequency of last word in the target phrase was excluded to reduce collinearity among predictors. *SE* = standard error. NS = native English speaker.

*worry*) corresponded to a 5% decrease in the production durations of the target segment (e.g., *don't have to*). In light of usage-based approaches (e.g., Bybee, 2010), this could mean that the higher frequency, and thus the stronger relation, between the last word in a target segment (e.g., *to*) and the last word in a phrase (e.g., *worry*) made the participants produce the last word in the target segment faster, leading to shorter production durations of the target segment.

Table 7 shows results for the low cutoff bin. Similarly, the coefficient for phrase frequency was negative, but phrase frequency effects did not reach significance, $\chi^2$ (1) = 3.21, $p$ = .073. There were significant main effects of participant group, $\chi^2$ (1) = 35.96, $p$ < .001, block order, $\chi^2$ (1) = 12.44, $p$ < .001, and the number of syllables in the target segments, $\chi^2$ (1) = 122.84, $p$ < .001. That is, the NNSs were

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

on average 17% slower that the NSs. Moreover, participants were on average 1% faster when they did the second experiment block, and a 1 *SD* increase in the number of syllables in the target segments led to about a 15% increase in production durations. The interaction between phrase frequency and block order was significant, $\chi^2$ (1) = 3.98, $p$ = .046. The negative coefficient indicates that, in the second block, participants produced the target segments in high-frequency phrases 2% faster than the target segments in low-frequency phrases. The interaction between phrase frequency and group, $\chi^2$ (1) = 0.39, $p$ = .527, was not significant, and neither were the frequencies of the last two words, $\chi^2$ (1) = 0.23, $p$ = .636, and the last three words, $\chi^2$ (1) = 0.97, $p$ = .325.

### *Discussion*

Research on frequency effects on L1 and L2 production has been limited compared to research on phrase comprehension. In particular, the study by Ellis et al. (2008) seems to be the only relevant study on compositional phrases beyond two words in NNSs. In the current experiment and in light of usage-based approaches (e.g., Bybee, 2010), NSs' shorter production durations compared to those from NNSs suggested that the target phrases were more entrenched in the NSs' linguistic representation, leading to faster productive processing. However, the results from both participant groups did not provide strong support for phrase frequency effects; participants processed high-frequency phrases faster than low-frequency phrases only in the second experiment block in the low cutoff bin. Given previous empirical studies, the absence of strong frequency effects may be attributable to the design and task instruction in the current experiment and the nature of the speech elicited.

The current experiment differed from previous related research on production durations of compositional phrases beyond two words in a few ways. Such previous research can be broadly divided into two groups: (a) four studies using an elicited production task (Arnon & Priva, 2013; Bannard & Matthews, 2008; Ellis et al., 2008; Tremblay & Tucker, 2011), and (b) two studies based on spontaneous speech (Arnon & Priva, 2013, 2014). Findings from the four studies in the first group have been inconclusive. Ellis et al. (2008) reported no frequency effects on production durations from adult NSs and NNSs, and Tremblay and Tucker (2011) similarly did not find frequency effects from adult NSs. In these two studies, as in the current experiment, participants were instructed to say the target phrases as fast as they could in an elicited production task. However, as pointed out, in Ellis et al.'s (2008) study, subpart frequencies were not controlled for, and the target phrases were of different types (i.e., complete or incomplete syntactic constituents). In contrast, in the current experiment, subpart frequencies were controlled for and the two phrases in each pair had the same constituency type. In addition, in Tremblay and Tucker's (2011) study, the regression analysis included many control variables, but the meaning of some of these control variables, including when significant (e.g., the interaction between frequency of the first word and the frequency of the third word in the target four-word phrases), is not obvious. Given these cross-study methodological differences, therefore, it might still not be safe to conclude that the results from the NSs in the current experiment and from these previous two studies are compatible.

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

The current experiment was perhaps more methodologically comparable to the elicited L1 production experiments by Arnon and Priva (2013) and Bannard and Matthew (2008). The former was conducted with adults and the latter with children aged 2–3 years old. Unlike the current experiment, these studies reported frequency effects on compositional four-word phrase production durations when subpart frequencies were controlled for. The incongruent results may have resulted from the remaining methodological differences. As in the current experiment, Arnon and Priva (2013) used a subset of the phrases from Arnon and Snider (2010) as stimuli. However, the researchers used a between-subject design to address a possible repetition effect resulting from a participant's reproduction of the identical first three words from a target pair (e.g., *don't have to worry* and *don't have to wait*). That is, the two phrases in each pair were assigned to two different lists, and each participant read only one of the lists. Therefore, one participant's production of *don't have to* in *don't have to worry* was compared against another participant's production of this same segment in *don't have to wait*. In their regression models, Arnon and Priva (2013) entered the average production durations of each participant (across all target stimuli) and the average production duration of each target segment (e.g., *don't have to*; across all participants) as control variables. The current experiment, however, used a within-subject counterbalanced design (i.e., every participant produced both phrases from each pair) because a between-subject design is less suitable for NNSs. That is, it is more difficult to match two different NNS groups on all variables known to affect L2 attainment, such as memory (e.g., Foster, Bolibaugh, & Kotula, 2014) and aptitude (DeKeyser, 2000). Thus, these methodological dissimilarities may have contributed to the incongruent results. Moreover, as in Arnon and Priva's (2013) study, the current experiment asked participants to wait until each target phrase disappeared before saying the phrase out loud. During the time the participants waited, there may have been a great deal of processing that was not captured. Given the relatively small amount of research on frequency effects on phrase production durations, these methodological issues could be investigated in future research.

Regarding Bannard and Matthew's (2008) study, one important difference between that study and the current experiment was the direction in the task. Bannard and Matthew (2008) asked children to "say the same thing" (p. 44) after hearing each target phrase from an audio clip. By contrast, in the current experiment, participants were instructed to say the phrase as fast as they could while still being accurate after reading each phrase on a computer screen. Possibly, the instruction in the current experiment prompted the participants to be more focused on producing the phrases; therefore, the difference between the production durations of high- and low-frequency phrases was less pronounced. In addition, in Bannard and Matthew's (2008) study, if children did not respond within a reasonable amount of time, the experimenter prompted them to respond once (e.g., by saying *Can you say that?*). Participants in the current experiment, by contrast, had only one chance to respond within the given time limit. In addition, the children may not have been as attentive as the adults in the current experiment. Bannard and Matthew (2008) asked each child to pronounce 32 phrases, 1 at a time, and retained only error-free productions in the analysis. The researchers later excluded a great deal of data due to production errors. In contrast, the mean production accuracy in the current experiment was

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

almost 100%. Possibly, the dissimilar instructions and the different amount of attention may have contributed to the incongruent results.

## General discussion

As Arnon (2015) pointed out, "[f]requency effects are not interesting in and of themselves. They are interesting because they reveal something about the [language] learning mechanisms ..." (p. 274). The current study investigated whether frequency effects can be observed in the processing of four-word English compositional processing in adult NSs and NNSs. Such effects are predicted in usage-based approaches, which attribute L1 and L2 acquisition to the interaction between domain-general human cognitive processes and input and predict that frequency affects processing of linguistic units at all levels (Bybee, 2010; Goldberg, 2006; Tomasello, 2009). In terms of receptive processing, the frequency effects observed from the NSs in Experiment 1, as well as from NSs in previous studies on four-word phrases (Arnon & Snider, 2010; Hernández et al., 2016), morphemes (Ambridge et al., 2015), single words (e.g., Diependaele et al., 2012), idioms (Nippold & Rudzinski, 1993), and two-word phrases (e.g., Gyllstad & Wolter, 2016), therefore provide empirical evidence supporting usage-based researchers' claim. Regarding NNSs, the results from Experiment 1 and recent findings from Hernández et al. (2016) extend empirical support for frequency effects on English L2 receptive processing of single words (e.g., Diependaele et al., 2013; Whitford & Titone, 2012) and two-word collocations (e.g., Gyllstad & Wolter, 2016; Sonbul, 2015; Wolter & Gyllstad, 2013) to longer compositional sequences. These likewise appear to corroborate the proposal in usage-based approaches to L2 acquisition (Ellis, 2011, 2013).

Given these findings, the next question is how to develop a corresponding L1 and L2 representation and processing psycholinguistic model. The results from Experiment 1 are compatible with usage-based models of language acquisition in which frequency of input affects language learning and construction representation at all levels (Bybee, 2010; Goldberg, 2006; Tomasello, 2009). As Arnon and Snider (2010) noted, an adequate model has to address several key issues, such as how an encounter of a specific lexical sequence is counted as an instance of multiple more abstract sequences, and the relationship between the representation of multiword phrases (e.g., *don't have to worry*), the subparts (e.g., *to worry*), and the more abstract linguistic units related to the subparts (e.g., an infinitive clause). Moreover, according to Ibbotson (2013), such a model needs to accommodate cognitive processes such as chunking and categorization, and should be able to expand and organize itself. Undoubtedly, a model for L2 speakers will be even more complicated due to various additional factors (e.g., L1 influence and L2 proficiency). Therefore, similar to the results from Hernández et al. (2016), the findings in Experiment 1 are in line with a unified connectionist model of L1 and L2 learning (MacWhinney, 2008), in which the processes underlying L1 and L2 acquisition share similarities and are not separable. In such models, both L1 and L2 acquisition rely on common cognitive processes, such as memory retention of word co-occurrences and chunking, and statistical information in input, including frequency, affects processing of L1 and L2 linguistics units at all levels, including compositional phrases.

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

Moreover, there does not need to be a qualitative difference between high- and low-frequency phrases; the entrenchment and processing difference is quantitative, resulting from different frequencies of previous phrase encounters.

Regarding phrase production, Experiment 2 in the current study lent weak support for phrase frequency effects from the adult NSs and NNSs, who demonstrated such effects in the receptive task in Experiment 1. To date, results from the relevant L1 research based on elicited phrase production durations, including the current study, have been mixed, and besides the current study, Ellis et al.'s (2008) study seems to be the only existing related study conducted with NNSs. As discussed, cross-study methodological differences could have contributed to the incongruent findings, and such differences should be investigated in future research, particularly given the relatively limited research on frequency effects on phrase production durations, especially in NNSs.

Perhaps it should also be pointed out that, unlike elicited production research, research based on four-word phrase production in NSs' spontaneous speech has more consistently reported frequency effects (Arnon & Priva, 2013, 2014). This consistency, together with frequency effects based on L1 spontaneous single-word production (e.g., Bell et al., 2003; Bybee & Scheibman, 1999), might suggest that the type of speech investigated (elicited vs. spontaneous) is another methodological factor contributing to whether phrase frequency effects can be observed in empirical research. If frequency effects in phrase production are related to "activation of multi-word lemmas" (Arnon & Priva, 2013, p. 366), in the current study it might be possible that by the time each target phrase disappeared from the computer screen in Experiment 2, the multiword lemmas had already been activated, so the difference in production durations between high- and low-frequency phrases was reduced. Possibly, compositional phrase frequency effects will be observed more consistently if spontaneous speech from NSs and NNSs is the subject of investigation. As in the case of comprehension, if there is more evidence for frequency effects on phrase production, such sensitivity will entail the need for a development of a psycholinguistic model that accommodates both word and multiword phrase frequency in speech production. As Arnon and Priva (2013) suggested, such a model can be an expanded version of the connectionist models of L1 production (e.g., Chang, 2002; Chang, Dell, & Bock, 2006). An adequate model must accommodate activation and competition among single words and multiword phrases during phrase production.

### *Limitations*

The current study contains some limitations that future research could address. As reviewers aptly pointed out, the outcome measure in Experiment 2, which was also used in previous studies on frequency effects on phrase production (e.g., Arnon & Priva, 2013; Bannard & Matthews, 2008; Tremblay & Tucker, 2011), is only one out of multiple possible measures of frequency effects. While usage-based approaches predict that phrase production durations should be frequency sensitive (e.g., Ellis, 2011, 2013) and the outcome measure in Experiment 2 allowed for a clear comparison of articulation time of the same three words in each target pair, this measure may not be the most sensitive measure. In particular, because the

instruction asked participants to wait until each phrase disappeared before saying each phrase aloud, there may have been a great deal of processing that was not captured by the outcome measure. In addition, despite an at least 2-day interval between the two experiments, possibly a repetition effect may still have existed; participants may have been primed after their previous exposure to the same phrases a few days earlier in Experiment 1. Furthermore, the break between the two experiment blocks may not have eliminated the possible repetition effects resulting from participants producing the same target segments across the two blocks. Such repetition/priming effects may have reduced frequency effects in the production task and might also be a possible reason why phrase frequency effects were observed only in the low cutoff bin in Experiment 2. Future production studies may therefore use other frequency sensitive measures, such as voice onset time or the duration between the onset of visual presentation of each phrase and the beginning of phrase production (e.g., Ellis et al., 2008; Jannsen & Barber, 2012), and address the other limitations in the design of the current study.

The current study also did not aim to explore the influence of L2 proficiency on phrase frequency effects. In a recent study, Wolter and Yamashita (2018) reported that NNSs demonstrate frequency effects to a greater extent when their English proficiency increases. This seems in line with Ellis's (2011, 2013) proposal that L2 speakers need to accumulate sufficient frequency information to exhibit frequency effects. However, Wolter and Yamashita focused only on two-word collocations, and therefore future studies could explore whether similar results can be obtained based on longer compositional phrases.[3]

With regard to the stimuli, because the target phrases constituted only a subset of English compositional phrases (i.e., four-word phrases with identical first three words in each pair), future studies can investigate if the results are generalizable to other phrases. Moreover, despite evidence that L2 phrases with a direct word-for-word L1 translation are processed faster than L2 phrases without such a translation (e.g., Wolter & Gyllstad, 2011, 2013; Wolter & Yamashita, 2018), the current study did not control for such an L1 influence. Finally, from a usage-based view, besides frequency, there are other types of statistical information in input that may play a role in language acquisition, such as mutual information, a measure of associative strength of constituent words in a phrase, and delta P, which indicates the probability of occurrence of a word when another word is present (Ellis, 2006a, 2006b; Gries, 2010, 2015). The role of such statistical information can therefore be a topic of further research.

## Conclusion

Motivated by usage-based approaches to L1 and L2 acquisition, the current study found frequency effects on receptive compositional English four-word phrase processing by both adult NSs and NNSs, lending support to the prediction in such approaches that these speakers can retain memory of word co-occurrences and that compositional phrase processing reflects frequency of previous phrase encounters (e.g., Bybee, 2010; Ellis, 2011). However, the study did not find similarly strong evidence for frequency effects on elicited phrase production durations from both participant groups. The results from the existing relevant research based on phrase

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

production durations, including the current study, have been mixed and may have resulted from cross-study methodological differences. This possibility could therefore be investigated in future research. Given the recent rise in usage-based L2 research, more studies on frequency effects on English L2 compositional multiword sequences are needed to support the ontological status of these effects.

**Acknowledgments.** This manuscript is based on my dissertation project. I am indebted to Karthik Durvasula and Qian Luo for their help with the computer programs used to create the speech production experiment and analyze the results. I also thank Jens Schmidtke and Attakrit Leckcivilize for their valuable suggestions on statistical analyses. In addition, I am grateful to Charlene Polio, Aline Godfroid, Susan Gass, and Shawn Loewen for their useful suggestions on earlier manuscript drafts. The manuscript also benefitted greatly from the constructive, detailed comments from Annie Tremblay and two anonymous reviewers. Final responsibility for any errors is my own.

## Notes

**1.** I thank Inbal Arnon for sending me examples of the fillers used in Arnon and Snider's (2010) study.
**2.** The transformation was based on the base-10 logarithm.
**3.** As a reviewer pointed out, because the effect of practice on frequency follows the power law of practice (e.g., DeKeyser, 1997), another possibility may be that frequency effects on L2 tasks may diminish as L2 proficiency increases; the effects may be greater in early stages of L2 learning than at an advanced level where improvements approach the asymptote.

## References

**Ambridge, B., Kidd, E., & Rowland, C. F., & Theakston, A. L.** (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, **41**, 239–273.

**Ambridge, B., & Lieven, E. M.** (2011). *Child language acquisition: Contrasting theoretical approaches*. Cambridge: Cambridge University Press.

**Arnon, I.** (2015). What can frequency effects tell us about the building blocks and mechanisms of language learning? *Journal of Child Language*, **42**, 274–277.

**Arnon, I., & Priva, U. C.** (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, **56**, 349–371.

**Arnon, I., & Priva, U. C.** (2014). The changing effect of word and multiword frequency on phonetic duration for highly frequent sequences. *Mental Lexicon*, **9**, 377–400.

**Arnon, I., & Snider, N.** (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, **62**, 67–82.

**Aylett, M., & Turk, A.** (2004). The smooth signal redundancy hypothesis: A functional explanation for the relationship between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, **47**, 31–56.

**Baayen, R. H., Davidson, D. J., & Bates, D. M.** (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, **59**, 390–412.

**Bannard, C., & Matthews, D.** (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, **19**, 241–248.

**Bates, D. M.** (2010). *lme4: Mixed-effects modeling with R*. New York: Springer.

**Bates, D. M., Mächler, M., Bolker, B. M., & Walker, S. C.** (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48.

**Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D.** (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America*, **113**, 1001–1024.

**Bley-Vroman, R.** (2009). The evolving context of the Fundamental Difference Hypothesis. *Studies in Second Language Acquisition*, **31**, 175–198.

**Boersma, P., & Weenink, D.** (2010). Praat: Doing phonetics by computer (Version 6.0.09) [Software]. Retrieved from http://www.fon.hum.uva.nl/praat/

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

**Bybee, J.** (2010). *Language, usage and cognition*. New York: Cambridge University Press.

**Bybee, J., & Hopper, P.** (2001). *Frequency and the emergence of linguistic structure*. Amsterdam: Benjamins.

**Bybee, J., & Scheibman, J.** (1999). The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics*, **37**, 575–596.

**Cedrus Corporation**. (2006). SuperLab Pro (Version 4.5) [Computer software]. San Pedro, CA: Author.

**Chang, F.** (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, **26**, 609–651.

**Chang, F., Dell, G. S., & Bock, K.** (2006). Becoming syntactic. *Psychological Review*, **113**, 234–272.

**Chomsky, N.** (1995). *The minimalist program*. Cambridge, MA: MIT Press.

**Cieri, C., Miller, D., & Walker, K.** (2004). The Fisher corpus: A resource for the next generations of speech-to-text. *Paper presented at the Fourth International Conference on Language Resources and Evaluation*, Lisbon, May 26–28, 2004.

**Davies, M.** (2013). Google scholar and COCA-academic: Two very different approaches to examining academic English. *Journal of English for Academic Purposes*, **12**, 155–165.

**DeKeyser, R. M.** (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition*, **19**, 195–221.

**DeKeyser, R. M.** (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, **22**, 499–533.

**Diependaele, K., Lemhöfer, K., & Brysbaert, M.** (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology*, **66**, 843–863.

**Durrant, P., & Schmitt, N.** (2010). Adult learners' retention of collocations from exposure. *Second Language Research*, **26**, 163–188.

**Ellis, N. C.** (1996). Sequencing in SLA: Phonological memory, chunking and points of order. *Studies in Second Language Acquisition*, **18**, 91–126.

**Ellis, N. C.** (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, **24**, 143–188.

**Ellis, N. C.** (2006a). Language acquisition as rational contingency learning. *Applied Linguistics*, **27**, 1–24.

**Ellis, N. C.** (2006b). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, **27**, 164–194.

**Ellis, N. C.** (2011). Frequency-based accounts of SLA. In S. Gass &A. Mackey (Eds.), *Handbook of second language acquisition* (pp. 193–210). London: Routledge/Taylor Francis.

**Ellis, N. C.** (2013). Second language acquisition. In G. Trousdale &T. Hoffmann (Eds.), *Oxford handbook of construction grammar* (pp. 365–378). Oxford: Oxford University Press.

**Ellis, N. C.** (2019). Essentials of a theory of language cognition. *Modern Language Journal*, **103** (Suppl.), 39–60.

**Ellis, N. C., Römer, U., & O'Donnell, M. B.** (2016). *Language usage, acquisition, and processing: Cognitive and corpus investigations of construction grammar*. Malden, MA: Wiley-Blackwell.

**Ellis, N. C., Simpson-Vlach, R., & Maynard, C.** (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, **42**, 375–396.

**Fernández, B. G., & 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.

**Foster, P., Bolibaugh, C., & Kotula, A.** (2014). Knowledge of nativelike selections in a L2. *Studies in Second Language Acquisition*, **36**, 101–132.

**Fox, J.** (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Thousand Oaks, CA: Sage.

**Fox, J., & Weisberg, S.** (2011). *An R companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage.

**Godfrey, J. J., Holliman, E. C., & McDaniel, J.** (1992). SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (Vol. **1**, pp. 517–520). San Francisco: IEEE.

**Goldberg, A. E.** (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, **7**, 219–224.

**Goldberg, A. E.** (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

**Gries, S. T.** (2010). Useful statistics for corpus linguistics. In A. Sánchez, and M. Almela (Eds.), *A mosaic of corpus linguistics: Selected approaches* (pp. 269–291). Frankfurt, Germany: Peter Lang.

**Gries, S. T.** (2015). 50-something years of work on collocations. *International Journal of Corpus Linguistics, 18*, 137–165.

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

**Hernández, M., Costa, A., & Arnon, I.** (2016). More than words: Multiword frequency effects in non-native speakers. *Language, Cognition and Neuroscience*, **31**, 785–800.

**Hoey, M.** (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

**Ibbotson, P.** (2013). The scope of usage-based theory. *Frontier in Psychology*, **4**. doi: 10.3389/fpsyg.2013.00255

**Jannsen, N., & Barber, H. A.** (2012). Phrase frequency effects in production. *PLoS ONE*, **7**. doi: 10.1371/journal.pone.0033202

**Littlemore, J.** (2009). *Applying cognitive linguistics to second language learning and teaching*. Basingstoke, UK: Palgrave Macmillan.

**Loewen, S., & Plonsky, L.** (2016). *An A–Z of applied linguistics research methods*. New York: Palgrave Macmillan.

**MacWhinney, B.** (2008). A unified model. In P. Robinson &N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 341–371). New York: Routledge.

**Muñoz, C.** (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Applied Linguistics*, **29**, 578–596.

**Myers, R.** (1990). *Classical and modern regression with applications* (2nd ed.). Boston, MA: Duxbury.

**Nippold, M. A., & Rudzinski, M.** (1993). Familiarity and transparency in idiom explanation: A developmental study of children and adolescents. *Journal of Speech and Hearing Research*, **36**, 728–737.

**Ortega, L.** (2013). SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn. *Language Learning*, **63**, 1–24.

**Peirce, J. W.** (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, **162**, 8–13.

**Pinker, S.** (1999). *Words and rules: The ingredients of language*. New York: Basic Books.

**Prasada, S., & Pinker, S.** (1993). Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes*, **8**, 1–56.

**R Core Team**. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

**Reali, F., & Christiansen, M. H.** (2007). Word-chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly Journal of Experimental Psychology*, **60**, 161–170.

**Reddy, S., & Stanford, J.** (2015). Toward completely automated vowel extraction: Introducing DARLA. *Linguistics Vanguard*, **1**, 15–28.

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

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

**Sosa, A. V., & MacFarlane, J.** (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of*. *Brain and Language*, **83**, 227–236.

**Supasiraprapa, S.** (2018). Prototype effects in first and second language learners: The case of English transitive semantics. *Bilingualism: Language and Cognition*, **21**, 618–639.

**Tomasello, M.** (2009). The usage-based theory of language acquisition. In E. Bavin (Ed.), *Handbook of child language* (pp. 69–87). New York: Cambridge University Press.

**Tremblay, A., Derwing, B., Libben, G., & Westbury, C.** (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, **61**, 569–613.

**Tremblay, A., & Tucker, B. V.** (2011). The effects of n-gram probabilistic measures on the recognition and production of four-word sequences. *Mental Lexicon*, **6**, 302–324.

**Valsecchi, M., Kuänstler, V., Saage, S., White, B. J., Mukherjee, J., & Gegenfurtner, K. R.** (2013). Advantage in reading lexical bundles is reduced in non-native speakers. *Journal of Eye Movement Research*, **6**, 1–16.

**Whitford, V., & Titone, D.** (2012). Second-language experience modulates first- and second-language word frequency effects: Evidence from eye movement measures of natural paragraph reading. *Psychonomic Bulletin & Review*, **19**, 73–80.

**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. *Studies in Second Language Acquisition*, **35**, 451–482.

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

**Wonnacott, E.** (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language*, **65**, 1–14.

**Wray, A.** (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

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

## Appendix A The target phrases

The two tables below show the target pairs of phrases and their frequencies per million words. These tables are based on the tables in Appendix A in Arnon and Snider's (2010) study.

### *Phrases in the high cutoff bin*

| | Phrase | Frequency | | Phrase | Frequency |
|---|---|---|---|---|---|
| 1. | out of the house | 12.00 | 9. | don't have to worry | 20.35 |
| | out of the game | 0.80 | | don't have to wait | 2.00 |
| 2. | we have to talk | 12.60 | 10. | I have to say | 21.00 |
| | we have to say | 0.90 | | I have to see | 1.40 |
| 3. | a lot of places | 12.80 | 11. | all over the place | 27.05 |
| | a lot of days | 0.70 | | all over the city | 0.85 |
| 4. | I want to go | 12.80 | 12. | I have a lot | 33.75 |
| | I want to know | 3.90 | | I have a little | 11.25 |
| 5. | don't know how much | 16.90 | 13. | on the other hand | 36.70 |
| | don't know how many | 10.15 | | on the other end | 4.80 |
| 6. | It's kind of hard | 17.10 | 14. | how do you feel | 36.95 |
| | It's kind of funny | 9.05 | | how do you do | 6.60 |
| 7. | a lot of work | 19.25 | 15. | I don't know why | 47.85 |
| | a lot of years | 2.55 | | I don't know who | 11.60 |
| 8. | go to the doctor | 19.70 | 16. | where do you live | 53.15 |
| | go to the beach | 6.95 | | where do you work | 4.35 |

https://doi.org/10.1017/S0142716419000109 Published online by Cambridge University Press

***Phrases in the low cutoff bin***

| | Phrase | Frequency | | Phrase | Frequency |
|---|---|---|---|---|---|
| 1. | we have to wait | 1.85 | 7. | it was really funny | 3.90 |
| | we have to leave | 0.35 | | it was really big | 0.20 |
| 2. | going to come back | 1.85 | 8. | I want to say | 5.60 |
| | going to come down | 0.55 | | I want to sit | 0.35 |
| 3. | you like to read | 2.10 | 9. | a lot of rain | 6.00 |
| | you like to try | 0.15 | | a lot of blood | 0.25 |
| 4. | out of the car | 2.60 | 10. | I have a sister | 6.95 |
| | out of the box | 0.30 | | I have a game | 0.05 |
| 5. | I have to pay | 2.80 | 11. | have to be careful | 7.10 |
| | I have to play | 0.15 | | have to be quiet | 0.15 |
| 6. | don't have any money | 2.80 | 12. | we have to talk | 12.60 |
| | don't have any place | 0.45 | | we have to sit | 0.25 |

**Cite this article:** Supasiraprapa S. (2019). Frequency effects on first and second language compositional phrase comprehension and production. *Applied Psycholinguistics* **40**, 987–1017. https://doi.org/10.1017/S0142716419000109