Collocational Processing in L1 and L2: The Effects of Word Frequency, Collocational Frequency, and Association

Doğuş Öksüz [1, 2], Vaclav Brezina [1] and Patrick Rebuschat [1, 3]

[1] Lancaster University, [2] University of Leeds, [3] University of Tübingen

Author note: We are grateful to the anonymous reviewers and the editors for their insightful comments and suggestions on earlier drafts of this article. We also thank Aina Casaponsa for her valuable comments and Miranda Barty-Taylor and Emma McClaughlin for proofreading. We also would like to thank Belma Haznedar, Yasemin Bayyurt, Leyla Marti, and Özlem Pervan for their support in data collection. Our research was supported by the Economic and Social Research Council, UK (grants ES/K002155/1 and EP/P001559/1) and the LEAD Graduate School & Research Network (grant DFG-GSC1028), a project of the Excellence Initiative of the German federal and state governments. Correspondence concerning this article should be addressed to: Doğuş Öksüz, Department of Linguistics and English Language, Lancaster University LA1 4YL, United Kingdom, d.oksuz@lancaster.ac.uk

Abstract

This study investigated the effects of individual word frequency, collocational frequency and association on L1 and L2 collocational processing. An acceptability judgment task was administered to L1 and L2 speakers of English. Response times were analysed using mixed-effects modelling for three types of adjective-noun pairs: (1) high-frequency, (2) low-frequency and (3) baseline items. This study extends previous research by examining if the effects of individual word and collocation frequency counts differ for L1 and L2 speakers' processing of collocations. This study also compared to what extent L1 and L2 speakers' response times are affected by mutual information and log dice scores, which are corpus-derived association measures. Both groups of participants demonstrated sensitivity to both individual word and collocation frequency counts. However, there was a reduced effects of individual word frequency counts for processing high-frequency collocations compared to low-frequency collocations. Both groups of participants were similarly sensitive to the association measures used.

Keywords collocation; multiword sequences; collocational processing; advanced learners, association measures, mutual information, log dice

Introduction

There has been a growing interest in research dedicated to the processing and use of multiword sequences (MWS). Regarding language acquisition and processing specifically, the importance of MWS is highlighted by usage-based approaches to language acquisition that have been gaining prominence (Bannard, Lieven & Tomasello, 2009; Christiansen & Chater, 2016; Tomasello, 2003). Within usage-based approaches, linguistic productivity is seen as a gradually emerging process of storing and abstracting MWS (e.g. McCauley & Christiansen, 2017; Goldberg, 2006; Tomasello, 2003). Such perspectives view both single words and MWS as essential building blocks for language acquisition and processing (Christiansen & Chater, 2016; Goldberg, 2006; McCelland, 2010). These approaches have received considerable empirical support from corpus studies, which report that large numbers of MWS are used in both spoken and written language. For example, Jackendoff (1997) carried out a small-scale corpus analysis of utterances used in a TV show and showed that MWS are used as often as single words in daily language. DeCock, Granger, Leech and McEnery (1998) estimated that MWS constitute up to 50% of both written and spoken native-speaker discourses. In addition to corpus evidence, there is substantial psycholinguistic evidence that both children (Arnon & Clark, 2011; Bannard & Matthews, 2008) and adults (Arnon & Snider, 2010; Jolsvai, McCauley & Christiansen, 2013; Tremblay, Derwing, Libben & Westbury, 2011) are sensitive to MWS during comprehension and production tasks. Furthermore, both first language (L1) and second language (L2) speakers appear to process MWS faster than matched control phrases (Siyanova-Chanturia, Conklin & van Heuven, 2011; Wolter & Yamashita, 2018).

MWS include collocations, lexical bundles, binomials, and idioms. Despite their obvious similarities, they vary considerably in completeness, structure, length, and transparency of meaning. Many studies have found that idioms (e.g. *kick the bucket*) are processed faster than matched control phrases (e.g. Rommers, Dijkstra, Bastiaansen, 2013; Siyanova-Chanturia, Conklin & Schmitt, 2011; Vespignani, Canal, Molinaro, Fonda & Cacciari, 2010). The same effect holds for lexical bundles (e.g. *in the middle of the*), which are defined as sequences of three or four words that occur as wholes at least 10 times per million words (Biber, Johansson, Conrad & Finegan, 1999). Tremblay et al. (2011) showed a processing advantage for lexical bundles in a self-paced reading experiment. Similar results have been reported for binomials, which are phrases consisting of two content words of the same class, with a conjunction in between (e.g. *knife and fork*). Siyanova-Chanturia et al. (2011), using eye-tracking, found that the original form of a binomial is processed faster than its reversed form (e.g *fork and knife*) by both L1 and L2 speakers. These findings provide empirical evidence that MWS are processed faster than matched novel phrases, due to their phrasal frequency, and predictability. This is consonant with usage-based approaches to language acquisition (Barlow & Kemmer, 2000; Ellis, 2002; Christiansen & Chater, 2016; Tomasello, 2003). These approaches underscore that language is rich with various types of distributional information such as frequency, variability and co-occurrence probability and that the human mind is sensitive to such distributional information (Erickson & Thiessen, 2015). Regarding the processing and acquisition of MWS specifically, two types of statistical information play important role, namely frequency and association (Ellis & Gries, 2015; Yi, 2018).

A prominent type of MWS that has received special attention in psycholinguistics, corpus linguistics, and language education studies is collocations. Different approaches to operationalising the complex notion of collocations have been put forth (McEnery & Hardie, 2011, pp. 122-123). The two most widely known approaches are the ‘phraseological approach’, and the ‘distributional’ or ‘frequency-based approach’. The phraseological approach focuses on the semantic relationship between two or more words and the degree of noncompositionality of their meaning (Nesselhauf, 2005; Howarth, 1998). According to phraseological approach, collocations are not simply free combinations of semantically transparent words, but they follow some selectional restrictions (e.g. *'slash' one's wrist* rather than *'cut' one's wrist*). The frequency-based approach draws on quantitative evidence on word co-occurrence in corpora (Evert, 2008; Gablasova, Brezina, & McEnery, 2017; McEnery & Hardie, 2012; Paquot & Granger, 2012), from which collocations are extracted using frequency cutoff scores and collocational association measures (see Evert, 2008; Gablasova et al. 2017, for a review of association measures). In this study, we adopt a frequency-based approach because we are primarily concerned with the effects of single word frequency, collocational frequency, and collocational strength on their processing by L1 and L2 speakers.

Factors affecting collocational processing

An important question is whether high-frequency collocations are a psychological reality for L1 and L2 speakers. Siyanova and Schmitt (2008) explored collocational processing by L1 and L2 speakers of English using a variation of an acceptability judgment task, finding that participants responded to high-frequency collocations faster than non-collocations. Durrant and Doherty (2010) conducted lexical decision tasks with L1 speakers to investigate if high collocation frequency or semantic association between the collocates led to faster processing of adjective-noun collocations. They found a priming effect in the processing of very high-frequency collocations, even if the collocates were not semantically associated. Wolter and Gyllstad (2013) looked at congruency effects on collocational processing for L2 speakers, and collocational frequency effects for both L1 and L2 speakers of English. They showed that collocations are processed faster in an L2 if they are congruent (i.e., a translation equivalent exists in the participants' L1). Furthermore, both L1 and advanced L2 speakers were sensitive to collocation frequency as they responded faster to more frequent collocations than less frequent collocations. Wolter and Yamashita (2015) developed this idea by examining whether collocations that exist in participants' L1 (Japanese) but not in the L2 (English) are still facilitated when processing collocations translated into the L2. They found no facilitation effect.

More recently Wolter and Yamashita (2018) investigated the congruency effect for L2 speakers. In addition, they examined single-word and collocational frequency effects for L1 and L2 speakers' processing of adjective-noun collocations. Replicating previous findings, they found a processing advantage for congruent collocations for L2 speakers, and no facilitation effect for the L1-only collocations translated into the L2. They suggested that the age or order in which something is learned affects how deeply it becomes entrenched in the language system, helping to explain the discrepancy between processing congruent and incongruent collocations. More specifically, as the learner gains L2 experience, the transferred congruent collocations from L1 to L2 become more entrenched through repeated exposure, while the nontransferable incongruent collocations become less entrenched due to lack of reinforcement. They also found that both L1 English and advanced L2 groups' processing were affected by word-level frequency and collocational frequency simultaneously, showing that L2 learners with advanced proficiency and L1 speakers' processing was affected by frequency information at multiple levels of representation.

Eye-movement studies also looked at L1 and L2 collocational processing. For example, Sonbul (2015) conducted a study which included L1 and L2 English speakers' on-line (eye-movement), and off-line (rating) measures of collocational processing. She developed three types of adjective-noun pairs: high-frequency collocations (e.g. *fatal mistake*), low-frequency (e.g. *awful mistake*), and non-attested synonymous pairs (e.g. *extreme mistake*). She examined how collocational frequency affects processing, finding that both L1 and L2 speakers are sensitive to collocation frequency in early measures of eye-movements, but not late measures.

Thus she suggested that collocations are not entirely fixed phrases; when reading an unexpected word pair, readers initially need longer time to process the pair, but once they incorporate it into a more general adjective-noun schema, they were able to process non-attested phrases comparably fast. Vilkaite (2016) looked at adult L1 speakers' eye-movements to test if non-adjacent collocations (e.g. *provide some of the information*) facilitated processing like adjacent ones (e.g. *provide information*). She found that L1 speakers are sensitive to both; adjacent and non-adjacent collocations showed similar processing advantages regarding entire-phrase reading times. However, the final-word reading measures only showed a processing advantage for adjacent collocations.

Overall, studies on collocational processing confirm that there is a processing advantage for collocations due to their high frequency. However, only a few studies have looked at the effects of probabilistic relationships of collocations, known as strength of association (see Evert, 2008; Gablasova et al. 2017), also defined as word-to-word contingency statistics (Yi, 2018) or transition probabilities (see McDonald & Shillcock, 2003; McCauley & Christiansen, 2017). In one example, McDonald and Shillcock (2003) analysed L1 English speakers' eye-movements to identify how strength of verb-noun collocations measured by transitional probabilities affect their processing. They found that initial-fixation duration was significantly shorter for verb-noun collocations with high transitional probability (e.g. *avoid confusion*) than pairs with low transitional probability (e.g. *avoid discovery*). However, Frisson, Rayner and Pickering (2005) found that transitional probabilities had no significant effect on collocational processing if contextual predictability was controlled. Nevertheless, they argued that contextual predictability (measured by cloze tests) involves some aspects of transitional probabilities, so one cannot entirely dismiss their effects on language processing. Ellis, Simpson-Vlach and Maynard (2008) investigated the psychological reality of MWS in academic contexts (e.g. *a wide variety of*) using a series of comprehension and production tasks. They found that L1 speakers' processing is affected by mutual information scores (MI-scores), which is a corpus-based association measure highlighting the rare exclusivity of word combinations (see Gablasova et al. 2017). However, advanced L2 speakers' processing of MWS appear to be affected by their phrasal frequency. These findings are interesting but because of the limited sample size and lack of control over confounding variables (e.g. single word and collocation frequency) the findings are limited.

Some recent experimental and computational modelling studies also looked at the effects of collocational strength on processing. Yi (2018) examined L1 and advanced L2 learners' sensitivity to frequency and association of adjective-noun collocations, revealing that both groups were sensitive to both measures, using MI-scores. Furthermore, advanced L2 speakers' sensitivity to collocational frequency and association statistics was considerably stronger than that of L1 speakers. McCauley and Christiansen (2017) compared L1 and L2 learners' use of MWS, employing a large-scale corpus-based computational model. They found that L2 learners are significantly more sensitive to the phrasal frequency of MWS than their associations, measured by MI-scores. Due to these contrasting findings it remains unclear whether L2 speakers are sensitive to collocational strength, or whether the corpus-based association measures used (e.g. MI) directly affect the findings with regard to speakers' sensitivity to collocational strength.

Operationalising collocational strength: Reviewing corpus-based association measures

The corpus-based association measures used in psycholinguistic studies are likely to directly and significantly affect the findings and consequently their insights into language learning and processing (Gablasova et al. 2017). Although various studies with a corpus linguistic focus have made efforts to standardise the conflicting terminology (e.g. Ebeling & Hasselgård, 2015; Evert, 2008; Gablasova et al. 2017), the rationale behind the selection of the association measures in psycholinguistic studies is not always fully transparent and systematic. Despite the availability of many association measures (see for example Evert, 2005; Wiechmann, 2008; Peccina, 2009 for comprehensive overviews), so far the MI-score has been predominantly used in psycholinguistic research either to extract collocations (e.g. Vilkaite, 2016) or to investigate language users' sensitivity to collocational strength (e.g. Yi, 2018). Therefore, we firstly review those studies which have chosen the MI-score, their justifications for using it, and the mathematical reasoning behind the MI-score. We then review the Log Dice (LD) measure as an alternative to the MI used in this study. Finally, other possible measures are briefly discussed.

The MI-score is a field-standard measure for calculating collocational strength in psycholinguistic research (e.g. Ellis et al. 2008; McCauley and Christiansen, 2017; Wolter & Yamashita, 2015; Vilkaite, 2016; Yi, 2018). It is described variously as a measure of appropriateness (Siyanova & Schmitt, 2008), coherence (Ellis et al. 2008), and significant co-occurrence (Wolter & Yamashita, 2015). It operates on a binary logarithmic scale expressing the ratio between the collocation frequency and the frequency of the random co-occurrence of the two words in the collocation (Church & Hanks, 1990). The random co-occurrence is similar to the corpus being a box in which all words are written on small pieces of paper and the box is shaken thoroughly (Gablasova et al. 2017). The reliability of this random co-occurrence model as a baseline is questionable since it assumes no structural properties of language, which is by definition not accurate. It favours low-frequency word pairs, whose components are likely to be low-frequency themselves (Garner, Crossley & Kyle, 2019; 2020; Schmitt, 2012). The measure has also a tendency to assign inflated scores to low-frequency combinations (see Appendix S1 for the mathematical equations of the MI and LD measures in the Supporting Information online). Thus the value does not only indicate the exclusivity of collocations but also how infrequently they occur in corpora (see also Evert, 2008; Gablasova et al. 2017). We must therefore be careful not to automatically interpret larger MI-scores as indicators of more coherent word combinations, because the MI-score is not constructed to highlight coherence or semantic unity of word combinations. Another disadvantage is that it operates on a scale that does not have theoretical minimum and maximum values, preventing easy interpretion of MI-scores for collocations extracted from different corpora.

As an alternative measure, Gablasova et al. (2017) introduced the LD, which has not yet been used in psycholinguistic and corpus-based language learning research. The LD-score uses the harmonic mean of two proportions that express the tendency of two words to co-occur, - relative to the frequency of these words in the corpus (Evert, 2008; Smadja, McKeown, & Hatzivassiloglou, 1996). Therefore the LD-score highlights exclusive, but not necessarily rare combinations and does not rely on the shake-the-box, random distribution model of language since it does not include the expected frequency in its equation. As a standardised measure on a scale with a fixed maximum value of 14 the LD-score is easier to interpret than the MI-score. It is therefore possible to see how far the value of a particular combination is from the theoretical maximum value (Gablasova et al. 2017). Word pairs with a high LD-score (over 13) include *vice versa, and zig zag* in the British National Corpus (BNC) XML edition. In sum, the LD measure is preferable to the MI-score if researchers aim to look at the exclusivity of collocations without low-frequency bias (Gablasova et al. 2017).

In practical terms, MI and LD measures capture slightly different aspects of the collocational relationships. The MI-score highlights rare exclusivity, since it is negatively linked to frequency. In other words, it rewards lower frequency combinations, for which there is less evidence in the corpus (see also Gablasova et al. 2017; Evert, 2008). For instance, the combination *ceteris paribus* receives a lower MI-score (raw frequency=46, MI=21) than *jampa ndogrup* (raw frequency = 10, MI=23.2), according to the BNC XML edition. Although both combinations are exclusively associated, the former combination is considerably more frequent than the latter one. Importantly, the LD is an ideal measure since it highlights exclusivity between words in the collocation without favouring low-frequency combinations (Gablasova et al. 2017). Furthermore, LD scores are reliable across corpora and sub-corpora because the scores are not affected by corpus size. Even though this study focuses on L1 and L2 speakers' sensitivity to MI and LD measures which highlight the exclusivity of collocations, we also should be aware of alternative association measures that capture other dimensions of collocational association. For example, Delta P, arising out of associative learning theory, highlights directionality of collocational strength (Gries, 2013). It identifies whether the first word is more predictive of the second one or vice versa (see Garner et al. 2018; 2019 for applications of the Delta P measure in learner corpus research). Dispersion is another dimension of collocational association, which takes into account the distribution of the node and collocates in the corpus (Gries, 2008). Cohen's d, the commonly used measure of effect size (Cohen, 1988), can be utilised as an association measure to explore the distribution of collocates in different texts or subcorpora (Brezina, McEnery, Wattam, 2015). Other association measures include t-score, MI2, MI3, z-score etc. Due to space constraints, these measures are not discussed here (see Brezina 2018: 66-75; Gries, 2008; 2013; Evert 2005 for a detailed review of association measures).

The Current Study

The present study first examines the prominence of single-word and collocation frequency information for processing high- and low-frequency collocations. More specifically, we aimed to examine whether L1 and L2 speakers' sensitivity to collocation and single-word frequency counts differ when processing high- and low-frequency collocations. Secondly, we wanted to examine whether there is a difference between L1 and L2 English speakers' sensitivity to association of collocations in relation to the specific measures used. L1 and L2 speakers' sensitivity to collocational association has been previously investigated by a few studies, but they produced contrasting findings. As pointed out, the literature has yet to reach a consensus on the effect of collocational association on L1 and L2 speakers' processing (e.g. McDonald & Shillcock, 2003; Ellis et al. 2008; Yi, 2018). Furthermore, the possible effect of specific association measures used on speakers' sensitivity remains underexplored. Therefore, the present study aims to test whether speakers' sensitivity to collocational association depends on how assocications are operationalised. In this way it may be possible to assess the extent to which the specific association measure used in the previous studies affected their findings. The following research questions were explored for the study:

1. Is there a difference between L1 and advanced level L2 speakers' sensitivity to both word-level and collocation frequency information when processing collocations?
2. Is there a difference between L1 and L2 speakers' sensitivity to word-level frequency information when processing high- and low-frequency collocations?
3. Is there a difference between L1 and L2 speakers' sensitivity to strength of collocations as measured by MI and LD scores?

Based on our review of theoretical positions and empirical studies we predicted that both L1 and L2 speakers are sensitive to both single-word and collocation frequency information simultaneously (see Wolter & Yamashita, 2018). However, we also expected that the frequency of the collocations would cause a difference in the prominence of word-level and collocation-level frequency information for both groups of participants. More precisely, the effect of individual word frequency information is expected to be weaker for processing high-frequency collocations than low-frequency collocations because with increasing frequency the whole would gain prominence relative to the part (Arnon & Cohen Priva, 2014). Finally, we predicted that the specific collocational association measures used affect L2 speakers' sensitivity to collocatinal association.

Method

Participants

The participants were a group of L1 English (native-speakers of English, *n*=30) and a group of advanced level L2 learners of English (L1 Turkish, *n*=32). The L1 English group consisted of 24 undergraduate and 6 postgraduate students all from a university in the UK. The L2 English group consisted of 22 undergraduate and 10 postgraduate students, all from two universities in Turkey. The LexTALE[1], a test of vocabulary knowledge for advanced learners of English (Lemhöfer & Broersma, 2012), - was administered to assess L2 English learners' vocabulary knowledge as a proxy for general English proficiency. The validity of the LexTALE as a measure of English vocabulary knowledge and indicator of general English proficieincy was assessed in a large scale study (see Lemhöfer & Broersma, 2012). LexTALE scores were found to be substantially and significantly correlate with Oxford Quick Placement Test, which is used to group learners in seven levels linked to the Common European Network for proficiency levels, ranging from beginner to upper advanced. To identify the L2 learners with advanced-level vocabulary knowledge, a cut-off LexTALE score was determined. Following the LexTALE norms reported by Lemhöfer and Broersma (2012), a LexTALE score of 80.5% (corresponding to the Oxford Placement Test of 80%) was used as a cut-off score to recruit advanced level L2 users of English. On average, the L1 English group had significantly larger vocabulary size than the L2 group (90.82 vs 84.85, $t_{(56.072)}$=5.15, $p < 0.05$). Twenty one participants in the L2 group had lived in an English speaking country for longer than one month (full biographical data for the participants are provided in Table 1).

Table 1 Means (standard deviations) for participant background variables

Variable	L1	L2
Age (years)	20.58 (2.16)	24.43 (4.01)
Gender (m/f)	12/17	16/16
Dexterity (r/l/both)	28/2/0	28/3/1
Mean starting age of learning English	-	10.96 (4.04)
Self-rated English speaking (1-6)	-	5.28 (0.44)
Self-rated English listening (1-6)	-	5.62 (0.48)
Self-rated English reading (1-6)	-	5.65 (0.47)
Self-rated English writing (1-6)	-	5.53 (0.49)
LexTALE scores	90.82 (3.71)	84.85 (5.22)

Note. English proficiency self-ratings are based on 1-6 scale (1=beginner, 6=advanced). LexTALE=Lexical test for advanced English learners. One L1 English speaker did not indicate gender.

Materials

To address our research questions, we used an acceptability judgment task (Wolter & Gyllstad, 2013). A key asumption underlying the task is that we should expect to see slower response times (RTs) for low-frequency collocations in comparison to high-frequency collocations for both the L1 and L2 groups. With these assumptions in mind, a total of one hundred and twenty English adjective-noun combinations were extracted from the BNC XML edition. Adjective-noun combinations were preferred following the methodological choice of Wolter and Gyllstad (2013) because variability in determiners in verb-noun combinations (e.g. *make a mistake* vs *make progress*) introduces another confounding variable, whereas adjective-noun combinations allows for more control over the item consistency by not including determiners. The items fell into one of the three critical conditions: (1) high-frequency collocations (*n*=30), (2) low-frequency collocations (*n*=30), (3) non-collocational (baseline) items (*n*=60). The non-collocational items were used for establishing threshold RTs, for measuring the relative RTs for the items in conditions (1) and (2). Single word frequency counts of the adjectives and nouns, collocation frequency counts, LD, and MI scores of the items were obtained from the BNC XML edition. For this study, we prefered to use nonlemmatised frequency counts at both the single-word and collocation level. Although arguments have been put forth favouring either the use of lemmatised over nonlemmatised frequency, Durrant (2014) found “no clear differences” between the two forms for predicting L2 learners’ knowledge of collocations.

To be able to extract the items for the three critical conditions, we explored the scales of adjective-noun collocations’ raw frequencies, and LD-scores in the BNC XML. In order to determine the frequency and LD cut-off scores for high- and low-frequency collocations, we selected 10 noun node words from various raw frequency counts with a high frequency count of 121591 (e.g. *people*), and a low frequency count of 8961 (e.g. *officer*). Using the selected noun nodes, a total of 4718 two-word adjective-noun combinations were extracted from the CQPWEB tool (Hardie, 2012). To determine the cut-off frequency counts and LD-scores for high-low-frequency collocations, we closely looked at the distribution of collocations’ raw frequency counts and the range of LD-scores for the adjective-noun pairs with various raw frequency scores (≤100, 100-200, 200-300, 300-400, 400≤) in the BNC. Unsurprisingly the frequency counts of the adjective-noun combinations follow Zipf-like skewed distributions, with a small number of high-frequency collocations, and a very large number of low-frequency adjective-noun combinations (see Appendix S2 in the Supporting Information online for a table of the noun nodes, and visual illustrations of collocations’ frequency information). To measure collocations’ strength of associations in each frequency bands, LD measure was used because it is not negatively linked to frequency (Gablasova et al. 2017). Adjective-noun collocations with raw frequency counts of ≥300 and LD-scores of ≥7 were defined as high-frequency collocations. Adjective-noun collocations with raw frequency counts between 10 and 150 and LD-scores between 2 and 4, within a 3-3 window span were defined as low-frequency collocations.

To select high-frequency collocations, the nouns in the BNC word frequency list were checked for whether they collocate with an adjective in a way that meets the cut-off raw frequency and LD-scores for high-frequency collocations. An initial list of 36 collocations satisfied the selection criteria for high-frequency collocations. Four of the collocations in the list were discarded because they were incongruent with Turkish (e.g *supreme court*, *british library*), considering the empirical evidence that lexical congruency affects collocational processing in L2 (Wolter & Gyllstad, 2011, 2013). Since the main goals of this study is to investigate whether there is a difference in L1 and L2 speakers' sensitivity to single-word and collocation frequency information, including incongruent collocations would be a confounding variable. To identify congruent items, the following procedure was followed. Initially, the first author (a native-speaker of Turkish with a high command of English) translated English items to Turkish. Then the translations were checked against the Turkish National Corpus (TNC), a large, balanced and representative corpus for modern Turkish with a size of 50 million words. The translated items which occur frequently in the TNC were identified as congruent collocations and the items which were not found in the TNC were considered as incongruent. Cognates were a concern since they may elicit faster RTs (Lemhöfer et al. 2008). We therefore discarded collocations whose component words were Turkish. Nonetheless, we could not fully eradicate all potential cognates. The number of remaining potential cognates corresponded to 8.3 percent of all items. A list of 30 high-frequency English collocations remained. The mean LD score for all high-frequency collocations was 7.80, with a low score of 7.0 (for the items *dark hair* and *left hand*), and with a high score of 10.95 (for the item *prime minister*).

To select low-frequency collocations, the unused nouns in the BNC high-frequency collocations word list, were checked for whether they collocate with an adjective in a way that meets the cut-off raw frequency and LD-scores for low-frequency collocations. The selected low-frequency collocations had raw frequency counts of between 10 and 150, LD-scores of between 2 and 4 within a 3-3 collocation window span. As with the high-frequency collocations, the low-frequency collocations were also congruent with Turkish. None of the nouns and adjectives used for the items in the high-frequency collocations were used for the items in the low-frequency collocations. However, single words (both adjectives and nouns) in both types of items were closely matched for item length, operationalised as number of letters, and frequency. A list of 30 low-frequency collocations were extracted. The mean LD score for all the low-frequency collocations was 3.24, with a low score of 2.54 (for the item, *away game*), and with a high score of 3.91 (for the item, *vital information*). Concordance lines were checked to ensure that for each of the high-frequency and low-frequency collocations, adjectives modified the nouns. All single word and collocational frequency counts were log transformed using SUBTLEX Zipf scale (Van Heuven, Mandera, Keuleers, and Brysbaert, 2014).

The baseline items consisted of random combinations of the nouns used for the high-low-frequency collocations with adjectives that had not been used for the high-low-frequency collocations. On the one hand, repeating the same nouns in different conditions was an ideal way of ensuring that the single word length and frequency counts of the nouns in the collocational and baseline conditions were perfectly matched. On the other, this meant that each noun appeared in the task twice and this inevitably introduced another potential confounder in that participants saw nouns twice under different conditions, potentially lowering the activation thresholds. To address this, all items were presented to the participants in an individually randomised order. Thus, any advantage gained from a seeing word for a second time was evened out both within the individual participant's test and across all of the participants as whole. That is to say, we used each noun once in a collocation and once in a baseline item. Adjectives in the collocational and baseline conditions were closely matched for frequency and length. All combined nouns and adjectives used to construct the baseline items were checked against the BNC to make sure that there was no co-occurrence. If any co-occurrence was found in the BNC, the LD-scores were checked to make sure that they were negative values. If the combinations produced positive LD-scores, the process was repeated. We eventually obtained a list of 60 baseline items; however, given the very large size of the BNC, it was not possible to fully eradicate the positive LD scores. We therefore decided to retain two items with positive but very low LD scores. The mean LD score for all baseline items was -0.93, with a low score of -3.22 (for the item *dirty time*) and with a high score of 0.45 (for the item *clear trade*). The baseline items had a raw frequency counts of ≤10. The concordance lines were checked to make sure that they were idiosyncratic rather than meaningful co-occurrences (See Appendix S3 for the full list of items in the Supporting Information online. Following open science practices, the items are also available from htpps://osf.io/dxvak/).

Table 2 Mean (standard deviations) for the test item characteristics

Item type	High-frequency collocations	Low-frequency collocations	Non-collocations	Statistical comparison
Item length	10.86 (2.97)	11.1 (2.3)	11.1 (2.52)	*W*=401, *p*=.46
Adjective frequency	5.17 (0.31)	5.17 (0.42)	5.15 (0.24)	*W*=467.5, *p*=.79
Noun frequency	5.36 (0.29)	5.36 (0.21)	5.36 (0.25)	*W*=415.5, *p*=.60
Collocational frequency	4.03 (0.34)	2.7 (0.3)	1.18 (0.52)	*W*=891, p<.05
Log Dice scores	7.8 (0.82)	3.24 (0.39)	-0.93 (0.85)	*W*=900, p<.05

Procedure

The RTs for the items in the three critical conditions were assessed by means of acceptability judgments. This task was administered using PsychoPy software (Peirce, 2007). It requires participants to indicate whether or not the items are acceptable. It has most frequently been used with grammatical acceptability in which judgments are more straightforward. However, the vast majority of the adjective-noun combinations are mostly grammatical unless the word combinations indicate something that is highly unlikely (e.g. *old child*). Therefore, adjective-noun combinations can be perceived as acceptable if some flexibility is used in interpreting them. To avoid this obstacle, we followed the alternate phrasing used by Wolter & Gyllstad (2013); and asked participants to indicate whether or not the word combinations were commonly used in English. The exact instructions were as follows:

> In this experiment, you will be presented with 120 word combinations. Your task is to decide, as quickly and accurately as possible whether the word combinations are commonly used in English or not. For instance, the word combination *harsh words*, is a commonly used word combination in English, but *complex force* is not a commonly used word combination in English. Please press the "YES" button on the game pad if the word combination is commonly used, and "NO" button if it is not commonly used in English.

The presentation sequence is shown in Figure 1. Firstly, the eye fixation (#########) was presented for 250ms, and followed by a blank screen. After the blank screen, the item was presented in lowercase in Times News Roman 12 pt. The item remained on the screen either until the participants indicate their responses (via pressing a button) or after a 4000ms timeout. They answered YES by pressing the button corresponding to the forefinger of the dominant hand, NO by pressing the button corresponding to the forefinger of the nondominant hand (in line with Ferrand et al. 2010; Robert & Rico Duarte, 2016; Sato & Athanasopoulos, 2018;

Shatzman & Schiller, 2004). The acceptability judgment task began with a practice session to familiarise the participants with the task. The participants were allowed a short break after the practice session. Most participants completed this task in 5-6 minutes. Afterwards, both groups of participants were administered the LexTALE test as a proxy for general English proficiency. In addition, the L2 group was also asked to complete a questionnaire to self-rate their perceptions of their English proficiency in the four skills; speaking, listening, reading and writing (see Table 1 for L2 groups' average self-rating proficiency scores).

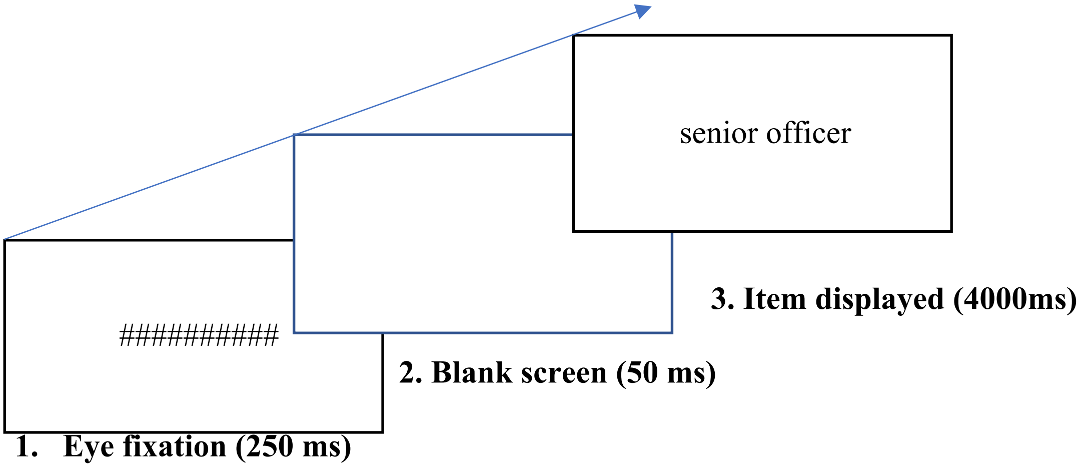

Figure 1 Presentation sequence for items in the acceptability judgment task.

Results

Preliminary Analyses

Following open science practices, all participant data including the LexTALE scores, RTs, is available from htpps://osf.io/dxvak/. The main concern of the present study was how the L1 and L2 participants processed the high- and low-frequency collocations they perceived as commonly used, compared to the baseline items they perceived as not commonly used. Therefore, we analysed the RTs to the high- and low-frequency collocations that received a "yes" response, and compared them to the baseline items that received a "no" response. This approach could have been potentially problematic in two ways. First if the majority of the high- and low-frequency collocations received a “no” response or a majority of the baseline items received a “yes” response. Fortunately neither was the case for both groups of participants. The L1 group judged 98% of the high-frequency collocations, and 78.11% of the low-frequency collocations to be commonly used in English, and they decided that 78.77% of the baseline items are not commonly used in English. The L2 group judged 97.5% of the high-frequency collocations, and 76.56% of the low-frequency collocations to be commonly used in English. They decided that 71.19% of the baseline items are not commonly used in English. The second reason this approach could have been problematic is that the corpus data we used do not fully represent the individual experiences of the participants (see also e.g., Durrant, 2013; González Fernández & Schmitt, 2015). That is to say, the individual differences in language experiences might have led some participants to judge some of the items based on their own language experiences of English which are different from the corpus-based evidence. However, considering the findings that both L1 and L2 speaker groups judged the vast majority of high- and low-frequency collocations as commonly used and baseline items as not commonly used, this was not the case for the present study. To begin the statistical analyses, we calculated mean RTs in milliseconds for each item type , that is the high- and low-frequency collocations that received “yes” responses and baseline items that received “no” responses. The mean RTs in the three conditions for both groups are also shown in Table 3 (see Appendix S4 in the Supporting Information online for a visual illustration of the same data).

Table 3 Response times in milliseconds (Standard deviations)

Item type	L1 (*n*=30)	95% CI	L2 (*n*=32)	95% CI
High-frequency coll.	892 (338)	[771.05-1012.95]	943 (383)	[810.3-1075.7]
Low-frequency coll.	1075 (431)	[920.77-1229.23]	1146 (477)	[980.73-1311.27]
Baseline items	1303 (527)	[1111.14-1491.58]	1326 (559)	[1132.32-1519.68]

Note: CI=Confidence interval

Model development

We used the Ime4 package (Bates, Maechler, Bolker, & Walker, 2015) in the R statistical platform (R Core Team, 2012) to construct mixed-effects models comparing RTs[2]. Before constructing the models, we prepared the data for analysis. The first step in this process was to prepare the RT data. Following the minimal data trimming choice by Gyllstad and Wolter, (2016), and Wolter and Yamashita (2018), only the responses that were faster than 450 ms, and the responses that timed out at 4.000 ms were excluded[3]. We carefully examined the histograms of log transformed and raw RT models' residuals. Since the distribution of the model residuals was not normal, the remaining RTs were log transformed (see also Baayen & Milin, 2010). The second step was to prepare the continuous predictors. All continuous predictors were centred and standardised, while the first versus second occurrence variables were treated as categorical. The third step was to recode the categorical factor group (L1 versus L2) using contrast coding. This provided some interpretational advantages for analysing the interactions. The recoding included converting the group factors into numeric variables (L1=0.5, L2=-0.5). The other categorical factor item type was coded using the treatment coding in which baseline items were defined as the reference level and high-frequency and low-frequency collocations were compared to the baseline items (baseline=0, high-frequency=1, low-frequency=1). Finally, the variance information factor scores (VIF-scores)[4] were calculated using the car package in R (Fox & Weisberg, 2019), to check whether there were any multicollinearity problems among the predictor variables. Finally, effect sizes of the models were computed using the MuMIn package in R[5] (Barton, 2019).

We constructed the first model to investigate whether there are significant group differences of overall mean RTs for any of the item types. It included participant and item as crossed random effects. We also had a by-subject (participant) random intercept for subject, a by-subject random slope for item type, a by-item random intercept for item, and also a by-item random slope for group. The following variables were included as fixed effects in the first model: group (L1 or L2), item type (high-frequency, low-frequency, or baseline), LexTALE scores, and the interaction between group and item type. Furthermore, we added item length, participants' age, gender, and the first versus second occurrence of the nouns (i.e. whether a participant was seeing a particular noun for the first or second time). The VIF-scores of item type, group, gender, age, LexTALE scores, and item length did not indicate any problems with multicollinearity (VIF-scores <2.00).

Table 4. Mixed effect model 1 (Comparing L1 and L2 speakers' RTs for high-frequency, low-frequency and baseline items)

Fixed effects	Estimate	*SE*	*t*	*P*
(Intercept)	0.500	0.377	1.32	.18
L1 vs L2	0.032	0.059	0.53	.59
Male vs Female	0.096	0.036	2.65	.01
Gender not stated vs Female	0.026	0.14	0.18	.85
Lextale scores	-0.004	0.003	-1.07	.28
Age	0.002	0.004	0.51	.60
High-frequency vs Baseline	-0.364	0.024	-14.87	<.0001
Low-frequency vs Baseline	-0.181	0.023	-7.74	<.0001
Item Length	0.034	0.007	4.67	<.0001
Order of occurrence	-0.005	0.007	-0.69	.48
L1 vs L2 x High-frequency (vs Baseline)	-0.021	0.038	-0.56	.57
L1 vs L2 x Low-frequency (vs Baseline)	-0.023	0.035	-0.67	.50

Note. R^2 marginal = 20. R^2 conditional = .42. SE = standard error. ImerTest[6] package (Kuznetsova, Brockhoff, Christensen, 2017) was used to compuete the *p*-values. One L1 English speaker did not indicate gender so that we included three levels (male, female, not stated).

The results revealed no significant differences between L1 and L2 groups either in terms of overall mean RTs (β= .032, [SE= .059], *p* = .59), or with respect to group by item type interactions. We ran a series of pairwise comparisons test to decompose the interactions between group and item type using the emmeans package in R with Tukey adjustments for multiple comparisons (Lenth, 2018). The results showed no significant differences between L1 and L2 groups' overall mean RTs for high-frequency (*Estimate* = .010., *z* = 0.22, *p* = .99), and low-frequency collocations (*Estimate* = .008., *z* = 0.15, *p* = 1). We also compared this model including the interactions (group by item type, LexTALE scores by item type) and without the interactions using a log-likelihood ratio test to find out whether the inclusion of these interactions produced a better-fitting model. There was not a significant difference between the two models according to the log-likelihood ratio test (chi-square = 0.46, *p* = .79), and this finding provided a further support to the conclusion that L1 and L2 speakers performed very similarly with respect to their RTs for all item types. As expected, both the high-frequency (β = -.346, [SE= .024], *p* < .0001) and the low-frequency collocations (β= -.182, [SE= .022], *p* < 0001) were responded to faster than the baseline items. Furthermore, relevelling the model to directly compare high-frequency collocations with low-frequency collocations revealed that high-frequency collocations were responded to faster than the low-frequency collocations (β= -.181, [SE= .023], *p* < 0001. Male participants had significantly slower RTs on average than female participants (β= .096, [SE= .036], *p* < .05), and the participants' ages do not seem to affect their RTs (β= .002, [SE= .004], *p* = .60). The effect of the LexTALE scores was not significant (β= -.004, [SE= .003], *p* = .28). Unsurprisingly, items with more letters received slower RTs (β= .034, [SE= .007], *p* < .0001). The effect of nouns' order of occurrence was not significant (β= -.005, [SE= .007], *p* = .48).

We constructed the second model to investigate the possible differences between L1 and L2 speakers' sensitivity to word-level frequency counts for adjectives and nouns and collocation frequency counts. For this model, we first eliminated the baseline items because nearly all of the baseline items had collocation frequency counts of zero. Furthermore, baseline items required a "no" response while high- and low-frequency collocations required a "yes" response in the acceptability judgment task. Considering the fact that different mechanisms might affect the processing of collocations' and baseline items' it is useful to analyse them separately. Because of the multicollinearity problem between collocation frequency and item type, we needed to discard the item type from this model (VIF=10.55, 9.78 respectively). As with the first model, this model also included participant and item as crossed random effects. Additionally, we included a by-subject (participant) random intercept for subject, a by-item random intercept for item, and a by-item random slope for group. In terms of fixed effects, the following variables were added to the second model: group (L1 or L2), single word frequency counts for adjectives and nouns, collocation frequency, and item length. Furthermore, group by adjective frequency, group by noun frequency and group by collocation frequency counts were added as interactions to the second model.

Table 5. Mixed effect model-2 (Investigating L1 and L2 speakers' sensitivity for adjective, noun and collocation frequency counts)

Fixed effects	Estimate	*SE*	*t*	*P*
(Intercept)	-0.059	0.020	-2.9	.004
Group (L1 vs L2)	-0.047	0.038	-1.23	.22
Adjective Frequency	0.023	0.009	2.42	.01
Noun Frequency	-0.000	0.009	-0.00	.99
Collocation frequency	-0.099	0.009	-10.31	<.0001
Item Length	0.04	0.009	4.23	<.0001
L1 vs L2 x Adjective Frequency	0.010	0.011	0.91	.36
L1 vs L2 x Noun Frequency	-0.000	0.012	-0.04	.96
L1 vs L2 x Collocation Frequency	-0.000	0.011	-0.08	.93

Note. R^2 marginal = 097. R^2 conditional = .30. SE = standard error. ImerTest package (Kuznetsova, Brockhoff, Christensen, 2017) was used to compuete the *p*-values.

As can be seen in Table 5, no significant differences of overall mean RTs were found between L1 and L2 groups (β= -.059, [SE= .038], p > 0.1). The results also yielded non-significant interaction effects between group and adjective frequency (*Estimate* = .010., z = 0.91 p = 0.35), group and noun frequency (*Estimate* = -0.00., z = -0.04, p = 0.96), group and collocation frequency counts (*Estimate* = -0.00., z = -0.08, p = 0.93). We compared this model with the main effects only version of the second model that excluded the interactions using a log-likelihood ratio test to find out whether the inclusion of the interactions produced a better-fitting model. There was not a significant difference between the two models according to the log-likelihood ratio test (chi-square = 0.86, p = 0.83), and this finding provided further support to the conclusion that L1 and advanced level L2 speakers were very similarly sensitive to the word-level and collocation frequency counts. As main effects, collocation frequency counts led to faster RTs (β= -.099, [SE= .009], p < .0001), while adjective frequency counts led to slower RTs (β= .033, [SE= .009], p < .05). The effect of noun frequency counts was not significant (β= -.000, [SE= .010], p = 0.9).

As pointed out above, due to the multicollinearity problem (between collocation frequency counts and item type) it was not possible to investigate the interaction between item type and single word frequency counts for adjectives and nouns in the second model. We therefore constructed the third model to explore possible difference in participants' sensitivity to word-level frequency information when processing high- and low-frequency collocations. We eliminated the collocation frequency from this model and added item type. For this model, the categorical factor item type were coded using the contrast coding scheme (High-frequency = .5, Low-frequency = -.5). We did not include the group either as a main effect or as an interaction between group and item type since their effects were not significant in the previous models. This model included participant and item as crossed random effects. We also included a by-subject (participant) random intercept for subject, a by-subject random slope for item type, a by-item random intercept for item. In terms of fixed effects, the following variables were added to the third model: item type (high-frequency, or low-frequency), word-level frequency counts for adjectives and nouns, item length, and interactions between word-level frequency counts for adjectives and nouns and item type.

Table 6. Mixed-effect model-3 (Investigating L1 and L2 speakers' sensitivity for single word frequency counts for processing high- and low-frequency collocations)

Fixed effects	Estimate	*SE*	*t*	*P*
(Intercept)	-0.047	0.021	-2.25	.02
High-freqeuncy vs Low-frequency	-0.184	0.020	-9.02	<.0001
Adjective frequency	0.008	0.010	0.85	.39
Noun frequency	-0.037	0.010	-3.49	<.0001
Item length	0.036	0.010	3.58	<.0001
Item type (High-freq. vs Low-freq.) x Noun frequency	0.045	0.022	2.0	.04

Note. R^2 marginal = 094. R^2 conditional = .30. SE = standard error. ImerTest package (Kuznetsova, Brockhoff, Christensen, 2017) was used to compuete the *p*-values (See Appendix S4 in the Supporting Information online for the confidence intervals alongside *p* values).

As shown in Table 6, high-frequency collocations were responded to faster than low-frequency collocations (β= -.184, [SE= .020], *p* < .0001). Noun frequency counts led to significantly faster RTs (β= -.037, [SE= .010], *p* < .0001). To interpret the interactions between noun frequency counts and item type, we first obtained the simple slopes for noun frequency counts by each level of item type (high-frquency vs low-frequency), using the emtrends function within the emmeans package in R (Lenth, 2018). There was a significant interaction between noun-frequency counts and item type (*Estimate* = 0.044., *z* = 2.0, *p* = .04), indicating that the participants' sensitivity to noun frequency counts varied depending on the frequency of the collocations. The interaction effect is shown in Figure 2. The effect of noun frequency counts on the participants' RTs were in the same direction for both high-frequency and low-frequency collocations. That is to say, as the noun frequency counts increased, participants'

RTs became faster. However, the effect of noun frequency counts on the participants' RTs for low-frequency collocations were stronger than the high-frequency collocations. More specifically, one unit of increase in noun frequency counts resulted in -0.059 log RT measure faster for low-frequency collocations, whereas one unit of increase in noun frequency counts resulted in -0.014 log RT measure faster for high-frequency collocations. The effect of adjective frequency counts was not significant (β= .008, [SE= .010], p > 0.3).

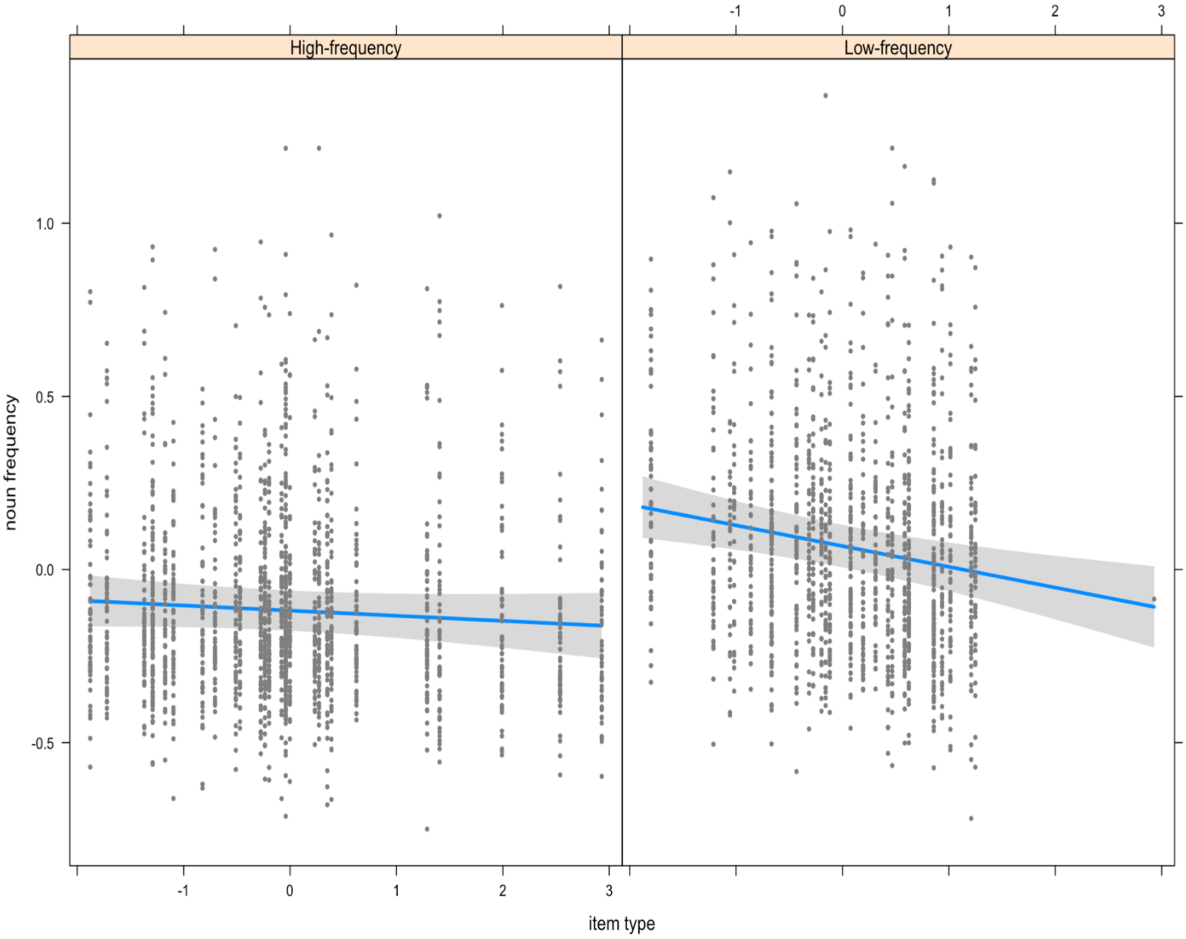

Figure 2. Interaction between item type and noun frequency counts

Finally, we constructed one more set of model that took into account the association statistics of collocations as measured by MI and LD-scores. We aimed to observe whether the way in which collocational association is operationalised would have an effect on L1 and L2 participants' sensitivity to them. We had two measures of collocational association (LD-score and MI-score). The high VIF scores of collocation frequency, LD-score, and MI-score (VIF=67.14, 186.75, 198.2 respectively) indicated a multicollinearity issue. In this case, we could not compare the co-efficients of the LD and MI measures in the same mixed model. Therefore, we decided to observe which association measure produce a better-fitting model of RT by comparing the Akaika Information Criteria values of the models.

We constructed two models, one includes LD-score, and the other includes MI-score. To observe whether there is a difference between L1 and L2 speakers' sensitivity to collocational strength as measured by MI and LD scores, we included the interaction between group and measures of association in the models. Then we compared the two models. The models included participant and item as crossed random effects. We included a by-subject random intercept for subject. We also had by-item random intercept for item, and also a by-item random slope for group. According to the AIC values, the model including LD-score is a better-fitting model (AIC=1500.2) than the model including MI-score (AIC=1511.7). Although the LD-score based model is a better-fitting model than the MI-score based one, the two models are not qualitatively different in their predictions (See Appendix S4 in the Supporting Information online for the tables of the two models including MI and LD scores).

Table 7 Mixed effects model 3 (Investigating L1 and L2 speakers' sensitivity to collocational strength as measured by Log Dice score)

Fixed effects	Estimate	*SE*	*t*	*P*
(Intercept)	-0.057	0.021	-2.65	.009
L1 vs L2	-0.046	0.038	-1.21	.22
LD-score	-0.093	0.012	-7.76	<.0001
Group (L1 vs L2) x LD-score	0.000	0.012	0.077	.93

Note. R^2 marginal = 076. R^2 conditional = .31. SE = standard error. ImerTest package (Kuznetsova, Brockhoff, Christensen, 2017) was used to compuete the *p*-values).

As can be seen in Table 7, results revealed no significant differences between L1 and L2 groups (*β*= -.046, [SE= .020], *p* > 0.2) in terms of mean RTs. As expected, LD-scores were associated with faster RTs (*β*= -.094, [SE= .009], *p* < .0001). The interaction between group and LD-score was not significant (*Estimate* = .0009., *z* = .07, *p* = .93).

Discussion

The results reveal that adult L1 and advanced L2 participants are sensitive to both word-level and collocation (phrasal) frequency information simultanously while processing two-word adjective-noun collocations. It should be noted that both noun frequency and collocation frequency counts led to faster RTs for both groups. The results further reveal that for both groups of participants, sensitivity to word-level frequency information in relation to nouns differs depending on the frequency of the collocations. More specifically, as the frequency of the adjective-noun collocations increase, the effect of noun frequency information becomes weaker for both L1 and L2 participants. This finding was expected because the increased use of collocations as two-word combinations is likely to make a difference in the prominence of individual word and collocation frequency information. Therefore, we see reduced effects of noun frequency and increased effects of collocation frequency information for high-frequency collocations. In the case of the effects of adjective frequency counts, findings suggest that they were associated with slower RTs for both groups. Finally, the results indicate that there was no difference in sensitivity to collocational strength between L1 and L2 groups irrespective of how they were operationalised. This finding was unexpected since MI and LD measures underlie different aspects of the collocational strength. We now focus on each of these findings in more detail.

In line with our hypothesis, the results of the mixed models 1 and 2 showed that the L1 and advanced L2 group's processing was affected by collocation frequency information while processing adjective-noun collocations. Both L1 and L2 groups responded to high-frequency collocations faster than the low-frequency collocations, and they also responded to low-frequency collocations faster than the baseline items. This indicates that both L1 and L2 groups needed a shorter time to process the collocations that occur more frequently. In addition, the results of the mixed model 2 indicated no difference between L1 and L2 participants' sensitivity to collocation frequency information. Therefore, the results of the present study add to the growing body of empirical evidence that both L1 and L2 speakers' processing is affected by phrasal frequency of MWS (e.g. Siyanova-Chanturia et al. 2011; Wolter & Yamashita, 2018; Yi, 2018) since both L1 and L2 groups' RTs became faster as collocation frequency increased. This is not to say, however, that participants' processing were only affected by collocation frequency information and their RTs were not affected by word-level frequency information. The results of the mixed model 2 show that noun frequency information led to significantly faster RTs. Furthermore, mixed model 2 indicated no difference between L1 and L2 participants' sensitivity to noun frequency information. Similar results have been reported in Wolter and Yamashita (2018), who also used an acceptability judgment task to compare an

L1 group and with two groups L2 speakers of differing proficiency for processing adjective-noun collocations. They found that all three groups' processing were affected by single word and collocation frequency information simultanously. In contrast, however they reported that the L2 groups appeared to rely more heavily on word-level frequency information than the L1 group.

Unlike Wolter and Yamashita (2018), L1 and L2 participants were comparably sensitive to word-level frequency information in the present study. The differences in findings is reconcilable, however. One possibility is that we have recruited a higher proficiency L2 group than they did. It is noteworthy that both the present study and Wolter and Yamashita (2018) found that L1 and L2 speakers' processing is affected by word-level and collocational frequency information simultaneously. These findings conflict with Wray's (2002, 2008) position that natives and non-natives process MWS in fundamentally different ways; that is to say, L1 speakers rely on their knowledge of meaning assigned to MWS whereas L2 speakers decompose MWS into individual words and rely heavily on the word-level information making up the MWS. On the contrary, the results of psycholinguistic research indicate that MWS are processed in a more unified way by L1 and proficient L2 speakers. For example, L1-based psycholinguistic and neurolinguistics studies have consistently reported that even if there is a processing advantage for frequent MWS as a whole, word-level frequency information still affects their processing, regardless of whether the phrases are idioms (e.g. Konopka & Bock, 2009; Snider & Arnon, 2012), complex prepositions (Molinaro, Canal, Vespignani, Pesciarelli, & Cacciari, 2013) or lexical bundles (Tremblay et al. 2011). In addition to the findings of the L1-based research, Wolter and Yamashita (2018) reported that both lower and higher proficiency L2 speakers are uniformly sensitive to to both word-level and collocation frequency information. The overall trend in the L2 speakers' RTs shows a progression from less reliance on word-level frequency to more reliance on collocation-level frequency with gains in proficiency. In the present study, with a very high proficiency group, we observed no significant differences between L1 and L2 speakers' reliance on word-level and collocation frequency information.

The findings that speakers are sensitive to both single word-level and phrasal frequency information also raises questions about how these different frequency measures interact when speakers process collocations on-line and whether there are differences between L1 and L2 speakers' reliance on word-level and collocational frequency information when they process high- and low-frequency collocations. Therefore, our second research question focussed on whether L1 and L2 speakers' reliance on word-level and collocation level frequency information differs depending on the frequency of the collocations. In line with our hypothesis, the results of the mixed model 3 indicated that the effect of noun frequency information on participants' RTs was stronger for the low-frequency collocations than the high-frequency collocations. In other words, for the high-frequency adjective-noun collocations, the effect of word–level frequency counts of the nouns on the RTs decreases, while the effect of collocation frequency increases. However, the results also showed that word-level frequency information still plays a role in the processing of even high-frequency collocations. On this point, the possible reasons for adjective frequency counts leading to slower RTs need to be addressed. As we failed to reliably establish an interaction effect between item type and adjective frequency counts, we need to apply caution in our approach to interpreting the findings. Nevertheless, it would be reasonable to suggest that when participants see collocations that include a very frequent adjective (e.g. *long time*), predicting the upcoming noun would be more difficult. This is also an expected finding from a corpus linguistics perspective because very high-frequency adjectives tend to form collocations with a wide range of nouns, but those collocations are unlikely to be highly-exclusive. The exclusivity of collocates refers to the extent to which the two words appear predominantly in each other's company (Gablasova et al. 2017). Exclusivity is strongly linked to predictability of co-occurrence when seeing one part of a collocation brings to mind the other part. Arguably, very high-frequency adjectives such as *long*, or *good* are unlikely to facilitate prediction because participants can not interpret them before they access to the nouns' meanings.

Similar patterns related to differing effects of single-word and multi-word frequency across the frequency continuum have been reported in Arnon and Cohen Priva (2014), focussing on L1 English speakers' phonetic duration in spontaneous speech. They found that the effect of multi-word frequency information increases with repeated usage while the effect of word-level frequency information decreases when producing high-frequency MWS. At this point, it is important to explore the usage-based notion of chunkedness, which positions the frequency and probability of input at the core of processing (Bybee & McCelland, 2005; Christiansen, & Chater, 2016; Ellis, 2002; Goldberg, 2006; Siyanova-Chanturia, 2015; Tomasello, 2003). They suggest that frequently used sequences become more accessible and more entrenched. Importantly this does not mean that frequently co-occurring MWS are stored and retrieved as unanalysed holistic units, which lack internal analysis, as Wray (2002, 2008) claims. Instead, usage-based approaches (e.g. Bybee, 2008, Ellis, 2002; Siyanova-Chanturia, 2015) suggest that frequently co-occurring MWS result in the growing prominence of the sequence relative to the parts - yet information related to the parts is still accessible. The present study (mirroring the findings of Arnon and Cohen Priva, 2014) provides empirical support to usage-based notions of chunkedness in two ways. First, participants' processing is affected by word-level frequency information for processing collocations, which suggests that collocations are not stored holistically. Second, the effect of the word-level frequency information of nouns differs depending on the frequency of the collocations. Furthermore, usage-based approaches to language acquision predict that the cumulative experience speakers have with a target language appears to similarly impact both L1 and L2 speakers (Ellis, 2002). The results of the present study and the study by Wolter & Yamashita (2018) provide evidence that L1 and L2 speakers processing is affected by word-level and collocation-frequency information.

Our third research question focussed on L1 and L2 speakers' sensitivity to strength of collocations as measured by MI and LD measures. Based on the previous literature (e.g. McCauley & Christiansen, 2017), we predicted that there would be differences between L1 and L2 participants' sensitivity to collocational strength. This prediction was not supported, as participants in both groups were similarly sensitive to the association statistic. It is possible to say that language users are sensitive to the strength of collocations irrespective of their identity as L1 and L2 speakers. Previous studies have produced conflicting results regarding L1 and L2 speakers' sensitivity to association statistics. For example, McCauley and Christiansen (2017) found that L2 learners' chunking scores improved in the raw frequency-based version of their computational model while L1 child and adult speakers' chunking performance improved in the MI-score based model. They concluded that there may be important differences between the way L1 and L2 speakers chunk and these differences cannot be explained only on the basis of amount of exposure. Yi (2018) found that L2 speakers were more sensitive to the MI-scores than L1 speakers. He concluded that language users are sensitive to the statistical regularities regardless of their identity as L1 and L2. The present study and Yi (2018) are comparable since both studies used a similar task with adjective-noun collocations, and with a fairly advanced group of L2 speakers. One possible reason for the differences in results between the two studies could be related to the fact that we have recruited a higher proficiency L2 group than Yi did. That is to say, as the level of L2 proficiency increases, L2 speakers' sensitivity to association statistics becomes more and more L1 like.

A further point for discussion is the LD-score in relation to the MI-score. According to the AIC values, the model including the LD-score is a better-fitting model than the MI-score one. This is not a surprising finding considering the features of the two measures. As Gablasova et al. (2017) observed, the LD measure is somewhat similar to the MI-score since it is designed to highlight exclusive word pairs. However, unlike the MI-score, it does not highlight rare exclusivity. In other words, the LD-score does not reward lower-frequency combinations. We can show the inconsistency of the MI-scores with an example from the high-frequency collocations used in the current study. One of the high-frequency collocations *social policy* (raw frequency=876, MI=3.74) receives a considerably lower MI-score than another high-frequency collocations *annual report* (raw frequency=641, MI=5.78). However, these high-frequency collocations *social policy* and *annual report* obtain fairly similar LD-scores (7.19 and 7.13 respectively). It is also important to note that the nouns *report* and *policy* have fairly similar raw frequency counts, however the adjective *social* (raw frequency=41649) occurs more frequently than the adjective *annual* (raw frequency = 8117) in the BNC XML edition. In this case, we can conclude that MI-score tends to highlight infrequent collocations whose, components ”may also be infrequent themselves” (Garner et al. 2018; Schmitt, 2012, p. 6).

Conclusion

The present study contributes to the growing body of research that both L1 and L2 speakers are sensitive to the frequency distributions of MWS at multiple grain sizes. More precisely, L1 and L2 speakers show sensitivity to both word-level and collocation frequency information simultanously while processing adjective-noun collocations. Furthermore, the effects of word-level and collocation level frequency information differ for processing low- and high-frequency collocations for both L1 and L2 speakers. As the frequency of the collocations increases, the effect of noun frequency information becomes weaker. It is possible to say that repeated usage of MWS leads to growing prominence of whole, but the part information is still accessible. Finally, there was no difference in sensitivity to association statistics between L1 and L2 groups irrespective of how they were operationalised. The findings of the present study are in line with the predictions of the usage-based approaches that the cumulative experience speakers have with a target language appears to similarly impact both L1 and L2 speakers (Ellis, 2002).

Although the present study sheds light on L1 and L2 speakers' sensitivity to frequency and association statistics while processing adjective-noun collocations, there are some limitations that need to be acknowledged. First, the acceptability judgment task used in this study may not be the most ideal one to examine the possible qualitative differences between L1 and L2 speakers' processing of MWS. This task is likely to require the participants to reflect on adjective-noun pairs and thus the RTs may indicate metalinguistic based processing rather than automatic (subconscious) processing. Second, the most of two-word adjective-noun collocations used in this study are likely to be considerably more frequent than the three-word sequences, which have been used in the some previous psycholinguistic studies (e.g. Arnon & Cohen Priva, 2014). Therefore, our findings are limited to two-word collocations and should not be generalised to other types of MWS. It should also be noted that in this study we sampled a highly proficient adult L2 population and we acknowledge that these findings may not apply to L2 populations at other proficiency levels or age groups. Another limitation of this study is that some of our items had cognates for Turkish. It would be ideal to fully eradicate them because they might be associated with faster processing (Lemhöfer et al. 2008).

In order to gain a more comprehensive understanding of the processing of MWS, future research needs to focus on L2 populations at different proficiency levels, and the individual differences among L1 and L2 speakers including both personal variables such as length of staying abroad and cognitive variables such as declerative memory. Furthermore, future research should also look at the processing of MWS other than collocations such as three or four-word lexical bundles to broaden the scope of the research. This research adopted frequency-based approach and drew on corpus evidence to identify collocations. However to reach a more complete picture of collocational processing, future research should focus on semantic relations between words (e.g. Gyllstad & Wolter, 2016) and L1 and L2 speakers' intuitons of semantic unity of collocations alongside their frequency counts. It is also crucial to acknowledge the importance of previous works at the intersection of experimental and corpus-based approaches to the use and processing of MWS. For example Rebuschat, Meurers, and McEnery (2017) brought together researchers in cognitive psychology, corpus linguistics, and developmental psychology. This type of multi-method approach is particularly useful for research in the processing and learning of MWS. The main reason is that corpora, as large databases, can provide direct information about language users' word selection and co-selection and reveal regularities in collocational patterns produced by L1 and L2 users which allows researchers to hypothesise about the factors involved in the acquisition, processing, and representation of collocations (Gablasova et al. 2017). As psycholinguists we should explore language users' sensitivity to various aspects of the distributional information including frequency, association, directionality and dispersion. Corpora can provide association measures that capture different dimensions of collocational relationships such as directionality (Delta P) and dispersion (Cohen's d). In future research we should critically evaluate the contribution of these association measures (Gablasova et al. 2017) and investigate language users' collocational processing through their lens. We should not be satisfied with one default option of association measure, no matter how popular.

Notes

1- We used the LexTALE test as a proxy for general English proficiency. It enabled us to to quickly and reliably identify learners with advanced knowledge of vocabulary. In a large scale validation study, LexTALE scores were found to be good predictors of vocabulary knowledge, and a fair indicator of general English proficiency (see Lemhöfer & Broersma, 2012).

2- Mixed-effect models were chosen because they allow for the inclusion of both participant and item as random effects. This enables the researchers to account the individual differences (e.g. slow versus fast RTs). It eliminated the need for separate analyses with participants and items (so called F1 and F2 analyses).

3- A total of 29 items were excluded from the analysis because they did not receive any response. The RTs shorter than 450 ms and longer than 4000 ms were also removed. Overall this accounted for less than 1% of the data (0.39%). Only 3 RTs were shorter than 450 ms and 27 items timed out at 4000 ms.

4- We used VIF-scores to detect strongly correlated variables in the mixed effects models, which tend to have unstable estimates and large standard errors (Levshina, 2015). As cut-off VIF-scores, researchers use different values, some of them strict such as 5 and others are less strict such as 10. To avoid any risk of multicollineairy, we used 5 as cut-off score.

5- The MuMIn package in R is used to compute the effect sizes of linear mixed-effects models. It produces two R^2 values for a fitted mixed effect model in two forms: marginal and conditional. Marginal R^2 values are only associated with fixed effects while conditional R^2 values are associated with both fixed and random effects.

6- We calculated the *p* values using the ImerTest package in the R statistical software (see Kuznetsova, Brockhoff, Christensen, 2015).

References

Aksan, Y., Aksan, M., Koltuksuz., A, Sezer, T., Mersinli, Ü., Demirhan, U., … Kurtoglu Ö. (2012). Construction of the Turkish National Corpus (TNC). *Proceedings of the 8th International Conference on Language Resources and Evaluation* (pp. 3223-3226). Istanbul, Turkey.

Arnon I., & Clark E, V. (2011). Why brush your teeth is better than – Children's word production is facilitated in familiar sentence – frames. *Language Learning and Development 7*(2), 107-129. https://doi.org/10.1080/15475441.2010.505489

Arnon, I., & Cohen Priva, U. (2014). Time and again: The changing effect of word and multiword frequency on phonetic duration for highly frequent sequences. *The Mental Lexicon 9*(3), 377-400. https://doi.org/10.1075/ml.9.3.01arn

Arnon I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*(1), 67-82. https://doi.org/10.1016/j.jml.2009.09.005

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science, 19*(3)*,* 241–248. https://doi.org/10.1111/j.1467-9280.2008.02075.x

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modelling children's early grammatical knowledge. *Proocedings of the National Academy of Sciences, 106*(41), 17284-17289

Barlow, M., & Kemmer, S. (Eds.). (2000). *Usage-based models of language.* Stanford, CA: The Center for the Study of Language and Information Publications

Barton, K. (2019). MuMIn: Multi-model inference. R package version 1.43.6. Retrieved from https://cran.r-project.org/web/packages/MuMIn/index.html

Baayen, R. H., & Milin, P. (2010). Analysing reaction times. *International Journal of Psychological Research* 3(2), 12-28. https://doi.org/10.21500/20112084.807

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

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English.* Harlow, England: Longman

Brezina, V. (2018). **Collocation Graphs and Networks: Selected Applications**. In P. Cantos-Gómez, & M. Almela-Sánchez (Eds.), *Lexical Collocation Analysis* (pp. 59-83). (Quantitative Methods in the Humanities and Social Sciences). Springer. https://doi.org/10.1007/978-3-319-92582-0_4

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: a new perspective on collocation networks. *International Journal of Corpus Linguistics 20(2)*, 139-173. https://doi.org/10.1075/ijcl.20.2.01bre

Bybee, J. (2008). Usage-based grammar and second-language acquisition. In P. Robinson & N. Ellis (Eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition*. New York: Routledge. 216-236.

Bybee, J. & McClelland, J. L. (2005) Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. In N. Ritter (Ed.), T*he role of linguistics in Cognitive Science.* Special Issue of The Linguistic Review, 22 (2-4), (pp. 381–410).

Christiansen, M., H. & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral & Brain Sciences, 39, e62.* https://doi.org/10.1017/S0140525X1500031X

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics, 16*(1), 22– 29. https://doi.org/10.3115/981623.981633

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciencies. Hillsdale, NJ: Lawrence Erlbaum Associates.

DeCock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learning English on computer* (pp. 67-79). London: Addison, Wesley, Longman.

Davies, M. (2004). BYU-BNC (Based on the British National Corpus from Oxford University Press). Available at http://corpus.byu.edu/bnc/

Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory, 6*, 125–155. https://doi.org/10.1515/CLLT.2010.06

Durrant, P. (2013). Formulaicity in an agglutinating language: The case of Turkish. *Corpus Linguistics and Linguistic Theory, 9*(1), 1–38. https://doi.org/10.1515/cllt-2013-0009

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*(2)*,* 143–188. https://doi.org/10.1017/S0272263102002024

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*(3)*,* 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental review, 37,* 66-108. https://doi.org/10.1016/j.dr.2015.05.002

Ebeling, S. O., & Hasselgård, H. (2015). Learner corpora and phraseology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 185–206). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Evert, S. 2005. The Statistics of Word Co-occurrences: Word Pairs and Collocations. Ph.D. thesis. Stuttgart: University of Stuttgart.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 1212–1248). Berlin, Germany: Mouton de Gruyter.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Meot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488-496. https://doi.org/10.3758/BRM.42.2.488

Fox, J. & Weisberg, S. (2019). *An {R} Companion to Applied Regression, Third Edition*. Thousands Oaks CA. Sage

Frisson, S., Rayner, K., & Pickering, M. J. (2005). Effects of Contextual Predictability and Transitional Probability on Eye Movements During Reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 862-877. https://doi.org/10.1037/0278-7393.31.5.862

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing and interpreting the evidence. *Language Learning, 67*(S1), 155–179. https://doi.org/10.1111/lang.12225

Garner, J.R., Crossley, S.A., & Kyle, K. (2018). Beginning and intermediate L2 writer's use of n-grams: An association measures study. *International Review of Applied Linguistics in Language Teaching. 58(1), 51-74* *https://doi.org/10.1515/iral-2017-0089*

Garner, J., Crossley, S. A., & Kyle, K. (2019). N-gram Measures and L2 Writing Proficiency. *System. 80* (1), 176-187. https://doi.org/10.1016/j.system.2018.12.001

Goldberg, A. (2006). *Constructions at work.* Oxford: Oxford University Press.

González Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have?: The effects of frequency and amount of exposure. *International Journal of Applied Linguistics, 166*(1), 94-126. https://doi.org/10.1075/itl.166.1.03fer

Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics, 13(4),* 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

Gries, S. Th. (2013). 50-something years of work on collocations: What is or should be next . *International Journal of Corpus Linguistics, 18,* 137–166. https://doi.org/10.1075/ijcl.18.1.09gri

Gries, S. T., & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Leaning 65, 228-255.* https://doi.org/10.1111/lang.12119

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

Hardie, A. (2012). CQPweb—Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics, 17*(3)*,* 380–409. https://doi.org/10.1075/ijcl.17.3.04har

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

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics, 19*(1), 24–44. https://doi.org/10.1093/applin/19.1.24

Jackendoff, R. (1997). *The architecture of the language faculty.* Cambridge, MA: MIT Press.

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017). An R 2 statistic for fixed effects in the generalized linear mixed model. Journal of Applied Statistics, *44*(6), 1086–1105. https://doi.org/10.1080/02664763.2016.1193725

Jolsvai, H. McCauley, S.M. & Christiansen, M.H. (2013). Meaning overrides frequency in idiomatic and compositional multiword chunks. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachdmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society,* (pp. 692-697). Austin, TX: Cognitive Science Society.

Konopka, A., & Bock, K. (2009). Lexical or syntactic control of sentence formulation? Structural general- izations from idiom production. *Cognitive Psychology, 58*(1), 68–101. https://doi.org/10.1016/j.cogpsych.2008.05.002

Kuznetsova A, Brockhoff PB, Christensen RHB (2017). “lmerTest Package: Tests in Linear Mixed Effects Models.” *Journal of Statistical Software*, **82**(13), 1– 26. https://doi.org/10185637/jss.v082.i13

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R.H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(1), 12-31. https://doi.org/10.1037/0278-7393.34.1.12

Lemhöfer, K. & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for for Advanced Learners of English. *Behavior Research Methods 44*(2), 325-343. https://doi.org/10.3758/s13428-011-046-0

Lenth, R. (2018). emmeans: Estimated marginal means, aka least-square means. R package version 1.2.4

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis.* Amsterdam, Netherlands: John Benjamins

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

McClelland, J. L. (2010). Emergence in Cognitive Science. *Topics in Cognitive Science, 2*(4), 751-770. https://doi.org/10.1111/tops.12258

McDonald, S. A., & Shillcock, R. C. (2003). Eye-movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science, 14*(6)*,* 648–652. https://doi.org/10.1046/j.0956-7976.2003.psci_1480.x

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice.* Cambridge, UK: Cambridge University Press.

Nesselhauf, N. (2005). Collocations in a learner corpus. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.14

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32,* 130–149. https://doi.org/10.1017/S0267190512000098

R Core Team. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Retrieved from: http://www.R-project.org

Rebuschat, P. Meurers, D., McEnery, T. (Eds.). (2017). Language learning research at the intersection of experimental, computational and corpus-based approaches: An introduction. *Special issue of Language learning, 67*(S1)*,* 6-13. https://doi.org/10.1111/lang.12243

Robert, C., Rico Duarte, L. (2016). Semantic Richness and Aging: The effect of number of features in the lexical decision task. Journal of Psycholinguistic Research, 45, 359-365. https://doi.org/10.1007/s10936-015-9352-8

Rommers, J., Dijkstra, T., & Bastiaansen, M. (2013). Context dependent semantic processing in the human brain: Evidence from idiom comprehension. *Journal of Cognitive Neuroscience 25*(5), 762-776. https://doi.org/10.1162/jocn_a_00337

Pecina , P. (2009). Lexical association measures and collocation extraction. *Language Resources and Evaluation, 44*(1-2), 137-158. https://doi.org/10.1007/s10579-009-9101-4

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: Experiments in behaviour made easy. *Behaviour Research Methods, 51*(1)*,* 195-203. https://doi.org/10.3758/s13428-018-01193-y

Sato, S., & Athanasopoulos, P. (2018). Grammatical gender affects gender perception: Evidence for the structural feedback hypothesis. *Cognition, 176,* 220-231. https://doi.org/10.1016/j.cognition.2018.03.014

Shatzman, K. B., & Schiller, N. O. (2004). The word frequency effect in picture naming: Contrasting two hypotheses using homonym pictures. *Brain and Language, 90(1-3),* 160-169. https://doi.org/10.1016/S0093-934X(03)00429-2

Smadja, F., McKeown, K. R., & Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics, 22,* 1–38.

Schmitt, N. (2012). Formulaic language and collocation. In C. Chapelle (Ed), *The encyclopedia of applied linguistics* (pp. 1-10). New York: Blackwell.

Snider, N., & Arnon, I. (2012). A unified lexicon and grammar? Compositional and non-compositional phrases in the lexicon. In S. Gries, & D. Divjak (Eds.), *Frequency effects in language* (pp. 127–163). Berlin, Germany: Mouton de Gruyter

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review, 64*(3), 429–458. https://doi.org/10.3138/cmlr.64.3.429

Siyanova-Chanturia, A., Conklin, K., Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research, 27*(2), 251–272. https://doi.org/10.1177/0267658310382068

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 776–784. https://doi.org/10.1037/a0022531

Siyanova-Chanturia, A. (2015). On the "holistic" nature of formulaic language. *Corpus Linguistics and Linguistic Theory, 11*(2), 285–301.

https://doi.org/10.1515/cllt-2014-0016

Sonbul, S. (2015). Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/on-line collocational processing. *Bilingualism: Language and Cognition, 18*(3), 419–437. https://doi.org/10.1017/S1366728914000674

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, MA: Harvard 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*(2)*,* 569-613. https://doi.org/10.1111/j.1467-9922.2010.00622.x

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology, 67*(6)*,* 1176-1190. https://doi.org/10.1080/17470218.2013.850521

Vespignani, F., Canal, P., Molinaro, N., Fonda, S., & Cacciari, C. (2010). Predictive mechanisms in idiom comprehension. *Journal of Cognitive Neuroscience, 22*(8), 1682–1700. https://doi.org/10.1162/jocn.2009.21293

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

Wiechmann, D. (2008). On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. Corpus Linguistics and Linguistic Theory, 4(2), 253-290. https://doi.org/10.1515/CLLT.2008.011

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

Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing: A comparison of congruent and incongruent collocations. S*tudies in Second Language Acquisition, 35*(3), 451–482. https://doi.org/10.1017/S0272263113000107

Wolter, B., & Yamashita, J. (2015). Processing collocations in asecond language. A case of first language activation? *Applied Psycholinguistics 36*(5), 1193-1221. https://doi.org/10.1017/S0142716414000113

Wolter, B., & Yamashita, J. (2018). Word frequency, collocational frequency, L1 congruency, and proficiency in L2 collocational processing: What accounts for L2 performance? *Studies in Second Language Acquisition, 40(2),* 395–416. https://doi.org/10.1017/S0272263117000237

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

Wray, A. (2008). Formulaic language: pushing the boundaries. Oxford: *Oxford University Press*

Yi, W. (2018). Statistical sensitivity, cognitive aptitudes, and processing of collocations. *Studies in Second Language Acquisition*, *40*(4)*,* 831-856. https://doi.org/10.1017/S0272263118000141

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Mathematical equations of the association measures, MI and Log Dice.

Appendix S2. Additional information for material development.

Appendix S3. Complete list of experimental items

Appendix S4. Additional figure and tables for data analysis.

Appendix S5. R code for statistical analyses