In press in *Language Learning* (as of August 16, 2022)

Title: Investigating L1 and L2 Speaker Intuitions of Phrasal Frequency and Association Strength of Multiword Sequences

Running head: Intuitions of Frequency & Association Strength

Authors & author affiliations

Wei Yi, Peking University

Kaiwen Man, University of Alabama

Ryo Maie, Michigan State University

Abstract

This study investigated the accuracy of L1 and L2 speakers' intuitions of phrasal frequency and association strength of collocations, as well as the linguistic influences that give rise to such estimations. L1 and L2 speakers of English judged 180 adjective-noun collocations as one of the following: high frequency, medium frequency, or low frequency, and high association, medium association, or low association. Results showed that neither L1 nor L2 speakers demonstrated accurate intuitions of phrasal frequency and association strength. Both groups of participants employed linguistic information at phrase and single-word levels when giving intuitive statistical estimates. Interestingly, judgments of phrasal frequency and association strength were found to be intertwined for both L1 and L2 speakers. Taken together, such findings shed new insight on our understanding of language users' statistical knowledge of multiword sequences.

Keywords: phrasal frequency, association strength, collocations, multiword expressions, statistical intuition

Acknowledgements

We would like to express our gratitude to Professor Michael Long for his support of this project. We gratefully acknowledge the insightful and constructive advice from Editor Dr. Emma Marsden, Associate Editor Dr. Scott Crossley and Dr. Kristopher Kyle, as well as the anonymous Language Learning reviewers throughout the review process.

Correspondence concerning this article should be addressed to Wei Yi, School of Chinese as a Second Language, Peking University, No.5 Yiheyuan Road, Haidian District, Beijing, 100871, China. E-mail: weiyisla@pku.edu.cn

Introduction

Judgments of frequency and probability of events have survival value for humans in daily activities. As part of human cognition, the use of language makes no exception in this regard. Natural languages are abundant in statistical regularities (Gries & Ellis, 2015). Over the past decades, the representation and processing of statistical information in language have received a considerable amount of attention. There is a growing body of evidence supporting the fact that L1 and L2 speakers demonstrate reliable statistical intuitions. However, research efforts have primarily focused on the accuracy of intuitive judgments of word frequency, leaving it unclear whether language users possess accurate intuition of other types of statistical information and for multiword sequences. More importantly, little is known about how language users come to such intuitive judgments (Alderson, 2007). To address these issues, the current study explored L1 and L2 speakers' intuitions of two kinds of statistical information of collocations, namely, phrasal frequency and association strength (i.e., the co-occurrence probability of words that constitute word sequences). To reveal the sources of information contributing to such intuitions, we also investigated influences of orthographic, phonological, and semantic characteristics of the words that constituted such larger-than-word units, along with corpus statistics at the phrase level.

Literature Review

Intuition of Frequency of Single Words

Frequency of occurrence is a fundamental piece of information people encode about their experience of language. It indicates how likely one encounters a linguistic unit and determines the degree of automaticity when a word or phrase is processed or retrieved (Gries & Ellis, 2015). Frequency effects have been studied extensively from various aspects. When it comes to lexical processing, ample evidence has shown that L1 and L2 speakers are sensitive to the frequency of single words (Diependaele, Lemhofer, & Brysbaert, 2013). Alongside this line of research, many studies have also examined language users' intuitive estimation of word frequency, born out of practical and theoretical concerns. Practically, psychologists need to estimate how often words occur in a language in order to investigate how word frequency affects lexical processing (Brysbaert & New, 2009). Meanwhile, applied linguists have to estimate word frequencies (especially when these are not available) so as to select materials that are worth teaching (McCrostie, 2007). Theoretically, investigations of word frequency intuition can contribute to models of human memory (Zacks & Hasher, 2002) and decision-making (Tversky, 1974). In the field of second language acquisition, many researchers (e.g., Ellis & Gries, 2015; Ellis, Romer, & O'Donnell, 2016) hold that language users are tuned to frequency of input, and knowledge of frequency and other probabilistic information can be acquired after decades of language use.

Research on language users' intuition of word frequency was initiated in the 1960s and 1970s (for a summary of studies on the intuition of word frequency, see Appendix S1 in the online supporting information). These studies typically follow the paradigm as below. First, objective frequency counts of preselected lexical items are extracted from corpora, which are assumed to reflect the language use of the society. Subsequently, participants are asked to judge how often the words are used in a language community or in personal experience, with their responses collected through the magnitude estimation method or the multiple rank order task (Shapiro, 1969). The former task requires participants to estimate how frequently a list of words occur by assigning numbers to each word based on certain response scales, whereas the latter asks participants to put the words into rank order according to their relative frequency. Finally, to evaluate the accuracy of language users' intuition of word frequency, the Pearson product-moment correlation (for magnitude estimation) or the Spearman rank-order correlation (for multiple rank order) between subjective frequency estimates and objective frequency counts is calculated.

Tryk (1968) sampled 100 nouns to represent the spectrum of the Thorndike-Lorge frequency counts and asked college students to estimate how often each word was used by average English speakers in daily conversation in a given period of time. Correlations between subjective frequency estimates and logged objective frequency counts were moderately high, ranging from .74 to .78. Shapiro (1969) selected 91 words from the Thorndike-Lorge (Thorndike & Lorge, 1944) and Kucera-Francis (Kucera & Francis, 1967) tabulations and had six groups of L1 English speakers respond in terms of word frequency either in spoken or written language. Participants first ranked the words within each block based on their intuition. They were then presented with the entire list of words and asked to assign numbers to the words in accordance with their relative frequency. Shapiro found high correlations between objective counts retrieved from corpora and participants' subjective estimates of word frequency, ranging from .92 to .98. Carroll (1971) borrowed 60 words from Shapiro (1969) and found that L1 speakers' estimation of word frequency highly correlated with objective counts. Backman (1976) took 50 words translated from Shapiro (1969) and had L1 speakers of Swedish rate in terms of frequency of occurrence in written language. Similarly, subjective frequency ratings were found to be highly correlated with objective frequency counts (r = .93).

L1 speakers' accurate intuition of word frequency was also reported in later studies. Ringeling (1984) requested L1 speakers of English and Dutch-English bilinguals to rank 24 nouns based on their perception of word frequency in the language and in personal experience. Participants showed reliable intuition of word frequency in the language, with the correlation between subjective frequency estimates and objective frequency of occurrence ranging from .74

to .90. In a norming study, Balota, Pilotti, and Cortese (2001) collected subjective frequency estimates for 2,938 monosyllabic English words from L1 speakers. Participants were required to rate each word in terms of how frequently it was encountered in general and in different domains. In line with previous studies, relatively high correlations between subjective frequency estimates and objective corpora frequencies were found (*r* = .78-.83). McCrostie (2007) further compared the intuition of word frequency of English instructors and college students. Both groups of participants were asked to arrange two lists of words in frequency order. Intriguingly, English teaching professionals' frequency judgments were no better than those of college students. Moreover, for both groups, their average judgment accuracy of words in the middle frequency range (*r* = .49-.51) was much lower than those in the whole frequency range (*r* = .83-.84).

Unlike the above-mentioned research, two studies (Alderson, 2007; Schmitt & Dunham, 1999) did not find evidence supporting the validity of language users' intuition of word frequency. Schmitt and Dunham (1999) selected 12 sets of near synonyms and required L1 and L2 speakers of English to assign frequency ratings relative to an anchor word within each lexical set. In contrast to previous work, they only found moderate correlations between subjective ratings of word frequency and corpus data (*r* = .53 and *r* = .58 for L1 and L2 speakers, respectively). Alderson (2007) conducted three experiments to investigate the accuracy of judgments of word frequency by professional linguists. The first experiment required participants to report how frequently 100 English verbs occurred in every million words, the second experiment asked participants to rank order 50 verbs according to their frequency, and the third experiment requested participants to rank order 25 verbs that were chosen to represent a more evenly distributed range of word frequency. Similarly to Schmitt and Dunham (1999), only moderate correlations were found between expert judgments of word frequency and objective frequency counts.

Intuitions of Phrasal Frequency and Association Strength of Multiword Sequences

Language consists of units of varying sizes, such as single words and multiword sequences. Multiword sequences are widely used by language users (Erman & Warren, 2000) and play a critical role in achieving native-like proficiency for L2 learners. For multiword sequences, two types of statistical information are crucial, namely, phrasal frequency and association strength (Gries & Ellis, 2015). Phrasal frequency indicates how often a word combination is encountered by language users, whereas association strength measures the co-occurrence probability of words that constitute the word combination, which is thought to be linked to language users' capability to predict the words following or preceding another word in a sequence (Gablasova, Brezina & McEnery, 2017). The conceptual distinction between phrasal frequency and association strength has been supported by recent empirical studies, showing that phrasal frequency alone is not adequate to explain language users' behavioral patterns (Gries & Ellis, 2015) and that association strength plays an important role in L1 and L2 speakers' online processing of multiword sequences, independently of phrasal frequency (e.g., Yi, 2018; Yi, Lu & Ma, 2017). Phrasal frequency and association strength do not necessarily go hand in hand—in fact, one can even expect highly frequent multiword sequences in which the constituent words are loosely associated (e.g., *short time*), and vice versa (e.g., *historic buildings*, for more examples, see Appendix S2).

Association strength of multiword sequences is measured by various metrics, including forward or backward transitional probability (e.g., McDonald & Shillcock, 2003), ΔP (e.g., Gries & Ellis, 2015), *t*-score (e.g., Gablasova et al., 2017; Wolter & Gyllstad, 2011), mutual information (MI, e.g., Yi, 2018; Yi et al., 2017), and log Dice (e.g., Gablasova et al., 2017; Öksüz, Brezina & Rebuschat, 2021). Overall, each of these association measures has its advantages and disadvantages. For instance, transitional probability and ΔP can identify the unidirectional association of constituent words within collocations, whereas *t*-score, MI, and log Dice assume that associations are mutual and quantify the strength of co-occurrence in both directions. Among the bidirectional association measures, *t*-score does not operate on a standardized scale and is not comparable across corpora (Hunston, 2002), whereas both MI and log Dice use a logarithmic scale and highlight the exclusivity between words in collocations. MI and log Dice are fairly similar to each other. However, they differ in a number of aspects: 1) MI expresses the ratio between the frequency of collocations and the frequency of random co-occurrence of the constituent words, whereas log Dice captures the tendency of two words to co-occur relative to the frequencies of these words in the corpus (Gablasova et al., 2017); 2) MI does not have a theoretical minimum and maximum, whereas log Dice has a maximum value (i.e., 14); 3) MI is said to reward rare word combinations (Gries & Ellis, 2015), whereas log Dice does not have such a bias (for instance, based on the British National Corpus, the highly frequent collocation “long time” has an MI value of 5.5 and a log Dice value of 9.4, yet for the low-frequency collocation “racial discrimination”, the MI and log Dice values are 11.8 and 10.0, respectively).

Given that single words and multiword sequences are both essential components of our mental lexicon and that L1 and L2 speakers are sensitive to statistical regularities underlying both types of linguistic units, it is natural to assume that statistical intuitions as reported for single words may extend to word combinations as well. So far, few studies have directly investigated language users’ intuition of phrasal frequency of multiword sequences. Backman (1978) instructed a group of L1 speakers of Swedish to estimate the phrasal frequency of eighteen three-word combinations. Similarly to Schmitt and Dunham (1999), he used a word sequence as the anchor and required the participants to assign frequency estimates relative to the anchor. The correlation between subjective and objective phrasal frequency was .56. Siyanova and Schmitt (2008) examined phrasal frequency judgments made by L1 and L2 speakers of English on 31 frequent (native-like) and 31 infrequent (learner) adjective-noun pairings, extracted from a learner corpus of writings. The frequent collocations were further divided into high and medium frequency bands, based on a cutoff point of frequency at 100 occurrences in the British National Corpus. Participants were requested to rate each collocation on a 6-point scale (1→6: *very uncommon*→*very common*). Overall, participants' ratings of phrasal frequency did not correlate highly with corpus-based data (*r* = .58 and *r* = .44 for L1 and L2 speakers, respectively). Interestingly, L1 speakers were found to be able to distinguish frequent collocations from infrequent collocations, as well as high-frequency collocations from medium-frequency collocations. By contrast, L2 speakers could only distinguish frequent from infrequent word combinations. Siyanova-Chanturia and Spina (2015) investigated L1 and L2 speakers' intuition of phrasal frequency of Italian collocations. They extracted 80 noun-adjective collocations from the Perugia Corpus and divided them into high-, medium-, and low-frequency bands. Additionally, they also created another group of noun-adjective combinations, which were incorporated as collocations that were of very low frequency. L1 and L2 speakers of Italian were instructed to report their intuition of phrasal frequency based on a four-point scale (i.e., high vs. medium vs. low vs. very low frequency). Results showed that both L1 and L2 speakers' judgments of collocation frequency were predicted by corpus frequency. Cohen's kappa was chosen to measure the agreement between language users' judgments of phrasal frequency and objective corpus-based frequency bands. Based on Cohen's kappa calculated individually for each item, they concluded that L1 and L2 speakers' intuitive judgments of phrasal frequency were highly accurate for collocations in the high-frequency band. Despite of the growing recognition of the importance of association strength for the representation and processing of multiword sequences, to the best of our knowledge, not a single study has examined language users' intuitive knowledge of the strength of association between the constituent words within multiword sequences.

Linguistic Influences on Intuitive Judgment of Statistical Regularities

Usage-based approaches hold that L1 and L2 speakers can acquire rich knowledge of linguistic units as their language experience accumulates (Ellis & Ogden, 2017). When it comes to the processing of multiword sequences, current evidence suggests that language users may access the knowledge of word combinations as well as their constituent parts (for a review, see Siyanova-Chanturia, 2015). For instance, L1 and L2 speakers have been found to be sensitive to constituent word and phrasal frequencies when processing collocations (e.g., Öksüz et al., 2021; Wolter & Yamashita, 2018). In addition to frequency, language users also encode phonological, orthographic, and semantic information of words. Following that knowledge of constituent words contributes to the processing of multiword sequences, one may expect that language users should make use of various lexical properties when intuitively judging word frequency, phrasal frequency, or association strength. Surprisingly, few studies have considered how orthographic, phonological, and semantic characteristics of words impact L1 and L2 speakers' statistical intuitions. Backman (1976) found that L1 speakers' subjective estimation of word frequency correlated with the pronounceability (i.e., the degree of difficulty to pronounce a word, *r* = .82) and comprehensibility (i.e., the degree of difficulty to comprehend a word, *r* = .65) of words. Using corpus frequency, orthographic neighborhood size (i.e., the number of words of the same length, generated by changing one letter), and meaningfulness as predictors for subjective estimates of word frequency, Balota et al. (2001) found that L1 speakers' intuition of word frequency was driven by objective corpus-based frequency as well as the meaningfulness of lexical items. Interestingly, neighborhood size also contributed to subjective frequency ratings, but only for highly familiar items. Siyanova-

Chanturia and Spina (2015) is the only study that investigated linguistic influences on L1 and L2 speakers' statistical intuition of multiword sequences. They incorporated the length (i.e., number of letters) and frequency of constituent words to predict intuitive judgments of phrasal frequency of noun-adjective Italian collocations. Their results showed that participants—especially L2 speakers—tended to assign higher frequency ratings to collocations that contained shorter nouns.

Research Questions

In sum, the current literature is limited in the following aspects. First, although there is ample research examining the accuracy of language users' intuition of frequency of occurrence, most studies have focused on single words, leaving it unclear whether similar patterns can be found for larger-than-word units. Second, no single study has explored language users' intuition of association strength of multiword sequences. Language users' knowledge of association strength and phrasal frequency are related (Yi, 2018; Yi et al., 2017). As acknowledged by Siyanova-Chanturia and Spina (2015, p.556), when judging phrasal frequency of multiword sequences, participants might also make use of their knowledge of association strength. Nevertheless, no research has been carried out to investigate whether knowledge of phrasal frequency contributes to judgment of association strength, and vice versa. Third, little research has been done to reveal the sources of information language users rely on when making subjective judgments about frequency and probability of language use, especially how knowledge of constituent words contributes to statistical intuitions of multiword sequences. Finally, studies that have examined L2 speakers' statistical intuition are relatively lacking. To bridge these gaps, the present study explored both L1 and L2 speakers' intuitions of phrasal frequency and association strength of collocations, while examining the contribution of various kinds of phrasal and lexical characteristics. Specifically, MI and log Dice were chosen as measures of association strength of collocations. As mentioned in the previous section, MI and log Dice capture the bidirectional relationship between constituent words in collocations and operate on normalized scales, which make our results comparable across corpora. Moreover, the choice of log Dice allowed us to examine the exclusivity of collocations without a low-frequency bias as in MI (Gablasova et al., 2017). In addition to corpus-retrieved word frequency, phrasal frequency, and association strength, we included word length (number of letters), phonological and orthographic neighborhood size, and concreteness (the extent to which a word refers to a perceptible entity) as predictors of statistical intuitions. Word length (Siyanova-Chanturia & Spina, 2015) and neighborhood size (Balota et al., 2001) were chosen because they have been found to moderate intuitions of frequency of words or multiword sequences, whereas concreteness was selected because it represents a fundamental semantic distinction among words and plays a key role in word recognition (Schwanenflugel, 1991). We asked the following research questions:

1. To what degree do L1 and L2 speakers' subjective judgements match corpus-retrieved phrasal frequency and association strength of collocations?
2. Are L1 and L2 speakers sensitive to corpus-retrieved phrase-level statistical information (i.e., collocation frequency, MI, log Dice) when intuitively judging the phrasal frequency and association strength of collocations?
3. How do orthographic, phonological, and semantic characteristics of constituent words contribute to L1 and L2 speakers' intuitive judgments of phrasal frequency and association strength of collocations?

Methodology

Participants

We recruited 194 participants, including 81 English learners (55 females) and 113 L1 English speakers (58 females). The L2 speakers were Chinese international students studying in U.S. colleges. The L1 speakers were residents in the U.S., and had earned at least a bachelor's degree at the time of data collection. On average, the L1 and L2 speakers were 34.1 (*SD* = 12.7) and 24.0 (*SD* = 3.5) years old, respectively. L2 speakers' average age of onset for learning English was 9.0 (*SD* = 2.7), and their mean length of residence in the U.S. was 30.7 months (*SD* = 23.5). Seventy-four L2 participants reported their most recent TOEFL iBT scores. Following the advice of an anonymous reviewer, L2 participants were classified either as intermediate (TOEFL score ≤ 94, *Min* = 70, *N* = 12) or advanced English speakers (TOEFL score ≥ 95, *Max* = 119, *N* = 62) based on their self-reported TOEFL total scores, following the TOEFL official guide.[1] Based on 5-point scales, L2 speakers' average self-reported English use outside the classroom was 3.4 (*SD* = 0.8), while their average rating of English proficiency was 3.3 (*SD* = 0.8), 3.2 (*SD* = 0.8), 2.9 (*SD* = 0.8), and 2.9 (*SD* = 0.8) for reading, listening, speaking, and writing, respectively (see Appendix S3 for more information about the L2 participants' characteristics).

Stimuli

One hundred eighty English adjective-noun collocations were borrowed from Yi (2018). They were retrieved from an online database, namely, Phrases in English (PIE: Fletcher, 2011). PIE was derived from the second edition of the British National Corpus (BNC), which is a balanced corpus consisting of 100 million words of modern British English, widely distributed in written and spoken domains. Following Wolter and Gyllstad (2013), collocations were defined as multiword sequences consisting of words that co-occur more frequently than would be predicted by chance, given the frequency of the constituent words. Specifically, Yi (2018) defined adjective-noun combinations as collocations if 1) they occurred at least once per million words in the BNC, and 2) the statistical association between the adjective and the noun, measured by MI, was higher than 3.0.

The collocations were sampled from the BNC such that they represented the whole range of frequency and association strength (MI) of adjective-noun combinations that met the above-mentioned criteria. Based on frequencies retrieved from the BNC, log Dice values were computed for each collocation. To ensure that the target colocations were familiar to L2 participants, five intermediate-to-advanced L2 speakers of English who did not participate in this study rated their familiarity with the collocations on a 5-point scale (1→5: *totally unknown*→*extremely familiar*). The average familiarity rating was 4.5 (*SD* = 0.4). To address the linguistic influences responsible for language users' intuitive judgments of phrasal frequency and association strength, lexical properties of the constituent words (i.e., word1 and word2), including word length, orthographic neighborhood size, and phonological neighborhood size, were retrieved from the CLEARPOND database (Marian, Bartolotti, Chabal, & Shook, 2012)—which provided an interface for obtaining phonological and orthographic neighborhood sizes across languages. Furthermore, concreteness ratings of the nouns within each collocation were borrowed from Brysbaert, Warriner, and Kuperman (2014). Language use might differ between British and American English. To evaluate this, phrasal frequency and association strength of collocations, as well as frequencies of the constituent words, were looked up in the Corpus of Contemporary American English (COCA), which is a balanced, large-scale corpus consisting of around one billion words. Overall, corpus data obtained from the BNC and the COCA were highly correlated (*r* = .73 for phrasal frequency, *r* = .78 for MI, *r* = .68 for log Dice, *r* = .90 for the frequency of the adjectives, and *r* = .87 for the frequency of the nouns). Frequencies were transformed to occurrences per million words before being transformed to their natural logarithm. Based on the BNC data, the collocations were then ranked from the lowest to the highest phrasal frequency and grouped into three groups (high vs. medium vs. low phrasal frequency), each containing 60 items. The same practice was adopted for association strength, resulting in the collocations being categorized into three bands (high vs. medium vs. low association strength). For the grouping of phrasal frequency and association strength, no borderline items existed (for a full list of the collocations, see Appendix S2). The mean phrasal frequency of collocations in the low, medium, and high band was 0.331 (*SD* = 0.173, range: 0.077-0.663), 1.284 (*SD* = 0.159, range: 1.072-1.896), and 2.426 (*SD* = 0.420, range: 1.984-3.780), respectively. On average, the association strength of collocations in the low, medium, and high band, measured by MI, was 5.108 (*SD* = 0.834, range: 3.366-6.157), 7.201 (*SD* = 0.601, range: 6.179-8.101), and 9.511 (*SD* = 1.199, range: 8.180-12.713), respectively. The average association strength of colocations, measured by log Dice, in the low, medium, and high band, was 6.855 (*SD* = 0.711, range: 4.763-7.761), 8.362 (*SD* = 0.316, range: 7.803-9.009), and 9.815 (*SD* = 0.648, range: 9.010-11.495), respectively (see Appendix S4 for characteristics of the selected collocations). Correlations among characteristics of the collocations can be seen in Appendix S5 in online supporting materials.

Instrument and Task

The collocations were incorporated into a questionnaire, which consisted of three sections. In the first section, participants answered several questions about their demographic information. In the second section, they were instructed to judge how frequently a collocation was used in English, based on a 3-point scale: low frequency, medium frequency, and high frequency. In the final section, participants were asked to estimate how strongly two words within a collocation were associated. The strength of association was explained to the participants as how likely the constituent words were able to predict the appearance of one word given the other, regardless of the direction of prediction. Participants were required to respond based on a 3-point scale: loose association (i.e., one word can hardly predict the other), medium association (i.e., one word can predict the other to some degree), and strong association (one word can strongly predict the other). Previous studies have mostly used the multiple rank order task or the magnitude estimation task, yet it would be rather unnatural to require participants to either rank order the collocations or assign numbers to them based on their self-evaluation of phrasal frequency and association strength. Instead, following the advice of Alderson (2007) and the practice of Siyanova-Chanturia and Spina (2015), participants received two forced-choice tasks, in which they had to judge the target collocations as one of the following: high phrasal frequency, medium phrasal frequency, or low phrasal frequency (for the frequency judgment task), and high association strength, medium association strength, or low association strength (for the association strength judgment task). For the judgment of phrasal frequency and association strength, the collocations were randomly grouped into nine blocks, each containing 20 collocations. Participants were presented with the target items with the order of the blocks as well as of items within each block randomized. Examples were given to help participants understand the tasks (for the entire questionnaire, see Appendix S6).

Procedure

The questionnaire was administered online through Qualtrics and Amazon Mechanical Turk. Qualtrics was used to deliver the questionnaire to international Chinese students studying at U.S. colleges. In contrast, Amazon Mechanical Turk was a better method to reach L1 speakers of English, considering that it has large pools of respondents. Moreover, Amazon Mechanical Turk allowed us to restrict the participants to be L1 English speakers living in the U.S. and holding a bachelor's degree. Two versions of the questionnaire were created, such that the order of the judgment of phrasal frequency and association strength of collocations was counterbalanced. Half of the participants received either version of the questionnaire. They were instructed that there was no time pressure, they should make judgments relying on their own intuition, and there was no right or wrong answer. The questionnaire took about 25 minutes.

Statistical Analysis

In this study, we followed the suggestion of Alderson (2007, p. 404) and computed judgment accuracies of phrasal frequency and association strength for L1 and L2 speakers, calculated as the proportion of participants whose intuitive ratings matched the corpus-based groupings (low vs. medium vs. high phrasal frequency, or low vs. medium vs. high association strength). To reveal how language users come to their subjective estimations, Bayesian mixed-effects multinomial models were separately run for phrasal frequency and association strength (measured by MI and log Dice). The seven L2 participants who did not report their TOEFL iBT scores were excluded from data analyses. For both sets of statistical models, the following predictors were included: proficiency (L1 speakers vs. advanced or intermediate L2 speakers), phrasal frequency band (low vs. medium vs. high), MI band (low vs. medium vs. high), log Dice band (low vs. medium vs. high), word1 frequency, word2 frequency, word1 length, word2 length, word1 orthographic neighborhood size, word2 orthographic neighborhood size, word1 phonological neighborhood size, word2 phonological neighborhood size, and word2 concreteness. To examine whether the influences of phrasal and lexical characteristics on language users' intuitive judgments of phrasal frequency and association strength differ among the three groups of speakers, interactions between proficiency and the other variables were added. Proficiency was dummy-coded such that intermediate and advanced L2 speakers were compared against L1 speakers. Similarly, phrasal frequency, MI band and log Dice band were dummy-coded, using the high band as the reference level. The incorporation of MI band and log Dice band as predictors when modeling subjective judgments of phrasal frequency enabled us to explore whether participants made use of statistical association information when judging the phrasal frequency of collocations. The same rationale applied to the inclusion of phrasal frequency band when modeling subjective judgments of association strength. Given that we already included orthographic, semantic, and phonological variables specific to each item for statistical analyses, only random intercepts of subjects were considered for both sets of models.

Bayesian estimation was utilized for modeling parameter estimation via the *MCMCglmm* package (Hadfield, 2010), which is housed in *R* (version 3.6.2, R Core Team, 2021). Convergence was assessed via potential scale reduction parameters[3] (Gelman et al., 2013), housed in the *coda* package (Plummer, Best, Cowles, & Vines, 2005). Two chains using 13,800 iterations with thinning of 4 to reduce autocorrelation among samplers were performed. Model parameter estimates, standard deviations, and their 95% credible intervals were summarized based on the posterior densities using the final 10,300 iterations after burn-in 3,500 (Gelman et al., 2013). For the current study, a potential scale reduction factor of 1.2 or less for each model parameter was used as the cut-off indicating convergence (Gelman et al., 2013). Default non-informative priors offered by the *MCMCglmm* package were used (Hadfield, 2010). Specifically, for the fixed effects, flat normal priors were specified with means of zeros, and large variances of 10^8. In addition, diffuse priors were used for estimating random effects by specifying two scalar parameters of the inverse Wishart prior as V = 1, and nu = 0.002, following the practice of Hadfield (2010). Bayesian model results are summarized by reporting the posterior expected values (means of posterior distributions) and 95% credible intervals of the parameters. For each parameter, if its 95% credible interval includes zero, then it is reasonable to infer that very likely the parameter (variable) can take a value near zero, indicating it does not explain much variation in the dependent variables (i.e., L1 and L2 speaker intuitions of phrasal frequency and association strength). We considered the effect of an independent variable to be reliable only when its credible interval does not contain zero. Full model results as well as trace plots for model parameters are available on Open Science Framework (https://osf.io/r9avk/?view_only=8d1986bb9a5f4acd85a35bb9b7efa064).

Results

Accuracy of L1 and L2 Speakers' Statistical Intuitions

Overall, the results suggested that for both L1 and L2 speakers of English, their subjective intuitions of phrasal frequency and association strength were not accurate. As can been seen from Figure 1, participants' judgment accuracy seemed to follow an increasing pattern as corpus-based phrasal frequency or association strength (whether measured by MI or log Dice) increased (except for intermediate L2 speakers' judgment of association strength for medium and high MI band collocations, or medium and high log Dice band collocations). L1 speakers' intuition of phrasal frequency seemed to be more accurate than that of L2 speakers, but only for low and medium band collocations. However, when it comes to intuitions of association strength, L1 speakers did not seem to have such an advantage over L2 speakers. Interestingly, advanced L2 speakers exhibited more accurate intuitions of high frequency or high association strength collocations than intermediate L2 speakers, yet such a pattern was reversed for low association strength collocations (for accuracies across the bands, see Appendix S7).

For each of the 180 collocations, we also calculated the proportion of participants whose ratings of phrasal frequency and association strength matched its corpus-based band (see Appendix

S8, S9, & S10 in online supporting materials for full data). The result is summarized in Table 1. For both phrasal frequency and association strength, only a small proportion of collocations received accurate subjective ratings from L1 and L2 speakers. Moreover, for both phrasal frequency and association strength, considerable variation was found across the target collocations and between L1 and L2 speakers. For instance, the high-frequency collocation "nuclear weapons" was perceived as being frequently used in English by most L1 (77.9%) and L2 speakers (74.2% and 66.7% for advanced and intermediate L2 speakers, respectively), whereas only 21.2% of L1 speakers, 14.5% of advanced L2 speakers and 16.7% of intermediate L2 speakers put the high-frequency collocation "hard work" into the high-frequency band. Similarly, the high-association-band collocation "civil war", measured either by MI or log Dice, was consistently rated as being strongly associated by L1 (66.4%) and L2 speakers (72.6% and 58.3% for advanced and intermediate L2 speakers, respectively), yet 25.7% of L1 speakers, 24.2% of advanced L2 speakers and 8.3% of intermediate L2 speakers labeled the high-association-band collocation "varying degrees" as a strongly associated word sequence. Such variations indicate that some extraneous factors other than phrasal frequency and association strength could have influenced the participants' subjective ratings.

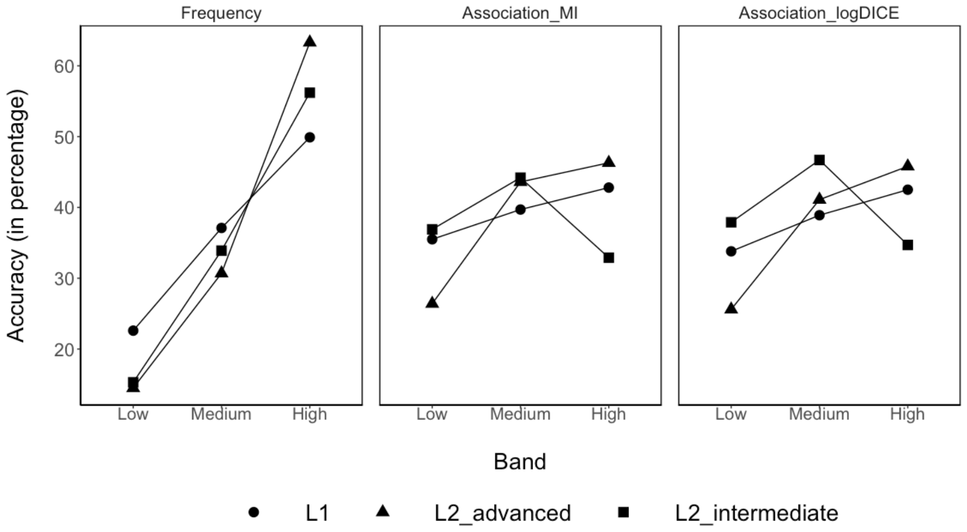

Figure 1 Accuracy (in percentage) of L1 and L2 speakers' intuitions of phrasal frequency (i.e., frequency) and association strength (i.e., association, measured by MI and log Dice).

Table 1 Summary of L1 and L2 speakers' intuitions of phrasal frequency and association strength for individual collocations (*k* = 180)

Statistical intuition	Proportion of matched responses for individual collocations			
		[0, 40%)	[40%, 60%)	[60%, 100%]
Phrasal frequency	L1 speakers	113	50	17
	L2 speakers (advanced)	117	28	35
	L2 speakers (intermediate)	101	60	19
Association strength (MI)	L1 speakers	101	69	10
	L2 speakers (advanced)	91	77	12
	L2 speakers (intermediate)	83	91	6
Association strength (log Dice)	L1 speakers	104	67	9
	L2 speakers (advanced)	74	99	7
	L2 speakers (intermediate)	105	65	10

Note. The values refer to the total number of collocations for which the proportion of participants' responses that matched the corpus-based groupings (i.e., low vs. medium vs. high) falls into each category. A parenthesis was used when the point or value was not included in the interval, whereas a bracket was used when the value was included.

Bayesian Model Results for the Judgment of Phrasal Frequency

Using MI as the measure of association strength, Bayesian mixed-effects multinomial modeling revealed that the model for the judgment of phrasal frequency included reliable effects of phrasal frequency band, MI band, interactions between proficiency and MI band, word1 frequency, word1 length, word1 orthographic neighborhood size, word2 frequency, word2 orthographic neighborhood size, word2 phonological neighborhood size, word2 concreteness, as well as interactions between proficiency and the following variables: word1 length, word1 phonological neighborhood size, word2 orthographic neighborhood size, and word2 concreteness (see Table 2). The effect of phrasal frequency band (*M* = -0.139, *SD* = 0.066, 95% CI = [-0.267, -0.008]) indicated that both L1 and L2 speakers were sensitive to corpus-retrieved phrasal frequencies when judging the phrasal frequency of collocations, with collocations in the low phrasal frequency band receiving lower ratings of phrasal frequency than word sequences in the high phrasal frequency band. The effect of MI band (*M* = -0.479, *SD* = 0.092, 95% CI = [-0.676, -0.311]), along with the interactions between proficiency and MI bands, showed that L1 and L2 speakers made use of MI band information when they judged the phrasal frequency of collocations. For both L1 and advanced L2 speakers, they assigned lower ratings of phrasal frequency to low-MI-band collocations than those labeled as high-MI-band sequences. However, for intermediate L2 speakers, reversed patterns were found—compared with high-MI-band word sequences, collocations in the low (*M* = -0.479+0.733) and medium (*M* = 0.614) MI bands received higher ratings of phrasal frequency.

The effects of word1 and word2 variables suggested that both L1 and L2 speakers' subjective estimations of phrasal frequency were affected by orthographic, phonological, and semantic properties of words that constituted the collocations. Specifically, target collocations containing higher-frequency adjectives (word1 frequency, *M* = 0.159, *SD* = 0.025, 95% CI = [0.111, 0.210]) and nouns (word2 frequency, *M* = 0.108, *SD* = 0.033, 95% CI = [0.041, 0.174]) were rated as being more frequently used than those with lower-frequency constituent words. The effect of word1 length (*M* = -0.130, *SD* = 0.014, 95% CI = [-0.157, -0.103]), along with the interaction between word1 length and proficiency (*M* = 0.062, *SD* = 0.028, 95% CI = [0.005, 0.114]), indicated that L1 and L2 speakers tended to rate collocations containing longer adjectives as being less frequently used than those consisting of shorter adjectives, even though such an effect among advanced L2 speakers was significantly weaker than that among L1 and intermediate L2 speakers. With respect to orthographic and phonological influences, complex patterns were found. Target collocations would be judged by L1 and L2 speakers as more frequent if they contained adjectives with more orthographic neighbors (word1 orthographic neighborhood size, *M* = 0.017, *SD* = 0.005, 95% CI = [0.009, 0.026]) and nouns with more phonological neighbors (word2 phonological neighborhood size, *M* = 0.015, *SD* = 0.003, 95% CI = [0.010, 0.021]). However, for word sequences consisting of nouns with more orthographic neighbors, L1 and advanced L2 speakers tended to judge them as being used less frequently (word2 orthographic neighborhood size, *M* = -0.032, *SD* = 0.006, 95% CI = [-0.043, -0.021]), whereas intermediate L2 speakers would judge them as being used more frequently (*M* = -0.032+0.049). Additionally, for advanced L2 speakers, they also tended to rate target collocations containing adjectives with more phonological neighbors as more frequent (L2-advanced: word1 phonological neighborhood size, *M* = 0.007, *SD* = 0.004, 95% CI = [0.000, 0.014]). With respect to the role of semantic characteristics, the effect of word2 concreteness (*M* = -0.115, *SD* = 0.022, 95% CI = [-0.155, -0.069]), along with the interaction between proficiency and word2 concreteness, suggested that target collocations consisting of more concrete nouns were perceived as being less frequent by L1 and advanced L2 speakers, yet such a pattern was reversed for intermediate L2 speakers (*M* = -0.115+0.285).

A separate Bayesian mixed-effects model was fitted for the judgment of phrasal frequency, using log Dice as the measure of association strength of collocations. Overall, results of this model (see Table 3) replicated the patterns reported when MI was used to measure association strength, which included reliable effects of association strength (log Dice band-low), word1 frequency, word1 length, word1 orthographic neighborhood size, word2 frequency, word2 orthographic neighborhood size, word2 phonological neighborhood size, word2 concreteness, as well as interactions between proficiency and word1 length /word2 orthographic neighborhood size/word2 concreteness. Nevertheless, some effects reported in the MI model (Table 2) were not replicated in this model; similarly, this model also revealed some effects that were not replicated in the MI model. Such inconsistent patterns may result from the fact that MI and log Dice capture different aspects of association strength (as mentioned in the literature review section) and suggest that the choice of association measures does have an impact on the results. Following this, the effects of phrasal frequency bands in the MI (i.e., phrasal frequency band-low) and log Dice models (i.e., phrasal frequency band-medium) might not lead to a strong conclusion that L1 and L2 speakers were sensitive to corpus-based phrasal frequencies when intuitively judging the phrasal frequency of collocations. Similarly, one might not confidently conclude that advanced L2 speakers would give lower ratings of phrasal frequency than L1 speakers, given that the effect of speaker (*M* = -1.220, *SD* = 0.552, 95% CI = [-2.307, -0.099]) was only found in the log Dice model. Lastly, the inconsistencies regarding the interactions between proficiency (L2-advanced) and word1 phonological neighborhood size/word1 frequency/word2 concreteness suggested that: 1) advanced L2 speakers might not necessarily be sensitive to the phonological neighborhood information of the adjectives that constituted the target collocations; 2) advanced L2 speakers might not differ from L1 speakers and intermediate L2 speakers in terms of the degree of sensitivity to the frequency of the adjectives that constituted the target collocations; 3) advanced L2 speakers might not differ from L1 speakers in terms of the degree of sensitivity to the concreteness of the nouns that constituted the target collocations.

Table 2 Bayesian mixed-effects multinomial modeling results for judgment of phrasal frequency (using MI as the measure of association strength)

Estimate	*M*	*SD*	95% CI
Frequency band-low	-0.139	0.066	[-0.267, -0.008]
MI band-low	-0.479	0.092	[-0.676, -0.311]
Proficiency (L2-intermediate): MI band-low	0.733	0.328	[0.111, 1.402]
Proficiency (L2-intermediate): MI band-medium	0.614	0.251	[0.177, 1.140]
Word1 frequency	0.159	0.025	[0.111, 0.210]
Word1 length	-0.130	0.014	[-0.157, -0.103]
Word1 ORTHO	0.017	0.005	[0.009, 0.026]
Word2 frequency	0.108	0.033	[0.041, 0.174]
Word2 ORTHO	-0.032	0.006	[-0.043, -0.021]

Word2 PHONO	0.015	0.003	[0.010, 0.021]
Word2 concreteness	-0.115	0.022	[-0.155, -0.069]
Proficiency (L2-advanced): Word1 length	0.062	0.028	[0.005, 0.114]
Proficiency (L2-advanced): Word1 PHONO	0.007	0.004	[0.000, 0.016]
Proficiency (L2-intermediate): Word2 ORTHO	0.049	0.023	[0.005, 0.094]
Proficiency (L2-intermediate): Word2 concreteness	0.285	0.086	[0.117, 0.446]

Note. Frequency band: band of phrasal frequency. ORTHO: orthographic neighborhood size. PHONO: phonological neighborhood size. Frequency band and MI band were dummy-coded, using the high band as the reference level. Proficiency was dummy-coded, with advanced and intermediate L2 speakers being compared against L1 speakers.

Table 3 Bayesian mixed-effects multinomial modeling results for judgment of phrasal frequency (using log Dice as the measure of association strength)

Estimate	*M*	*SD*	95% CI
Proficiency (L2-advanced)	-1.220	0.552	[-2.307, -0.099]
Frequency band-medium	0.154	0.077	[0.002, 0.314]
log Dice band-low	-0.522	0.080	[-0.696, -0.386]
Word1 frequency	0.163	0.025	[0.112, 0.209]
Word1 length	-0.135	0.015	[-0.164, -0.106]
Word1 ORTHO	0.011	0.005	[0.001, 0.020]
Word2 frequency	0.091	0.031	[0.032, 0.153]
Word2 ORTHO	-0.030	0.005	[-0.039, -0.019]
Word2 PHONO	0.016	0.003	[0.010, 0.020]
Word2 concreteness	-0.119	0.021	[-0.158, -0.077]
Proficiency (L2-advanced): Word1 frequency	0.097	0.045	[0.010, 0.189]
Proficiency (L2-advanced): Word1 length	0.058	0.026	[0.005, 0.105]
Proficiency (L2-intermediate): Word2 ORTHO	0.049	0.023	[0.005, 0.010]
Proficiency (L2-intermediate): Word2 concreteness	0.243	0.086	[0.079, 0.402]
Proficiency (L2-advanced): Word2 concreteness	0.083	0.040	[0.004, 0.164]

Note. Frequency band: band of phrasal frequency. ORTHO: orthographic neighborhood size. PHONO: phonological neighborhood size. Frequency band and log Dice band were dummy-coded, using the high band as the reference level. Proficiency was dummy-coded, with advanced and intermediate L2 speakers being compared against L1 speakers.

Bayesian Model Results for the Judgment of Association Strength

Using MI as the measure of association strength, Bayesian mixed-effects modeling revealed that the model for the judgment of association strength included reliable effects of MI band, phrasal frequency band, the interaction between proficiency and phrasal frequency band, word1 frequency, word1 length, word1 orthographic neighborhood size, word2 length, and the interaction between proficiency and word2 concreteness (see Table 4). The effect of MI band (*M* = -0.638, *SD* = 0.102, 95% CI = [-0.842, -0.437]) indicated that L1 speakers and L2 speakers (advanced or intermediate) were all sensitive to corpus-based association strength information when judging the degree of association of target collocations. The effects of phrasal frequency bands (low and medium), along with the interaction between proficiency (L2-intermediate) and phrasal frequency band (medium), suggested that L1 and L2 speakers also made use of corpus-based phrasal frequencies when judging the association strength of word combinations, with intermediate L2 speakers being more sensitive to such statistical information than L1 and advanced L2 speakers. Specifically, collocations in the low and medium phrasal frequency bands received lower ratings of association strength than those in the high phrasal frequency band. With respect to the influences of lexical characteristics of constituent words, the effect of word1 frequency (*M* = -0.266, *SD* = 0.027, 95% CI = [-0.322, -0.216]), word1 length (*M* = -0.042, *SD* = 0.016, 95% CI = [-0.073, -0.012]) and word1 orthographic neighborhood size (*M* = -0.014, *SD* = 0.005, 95% CI = [-0.024, -0.004]) suggested that target collocations containing adjectives that were more frequent, longer, and with more orthographic neighbors tended to be perceived as being associated less strongly. In addition, the effect of word2 length (*M* = 0.040, *SD* = 0.014, 95% CI = [0.014, 0.067]) indicated that target collocations containing longer nouns were rated as being associated more strongly than those with shorter nouns. Lastly, the interaction between proficiency (L2-advanced) and word2 concreteness (*M* = -0.089, *SD* = 0.042, 95% CI = [-0.017, -0.081]) showed that advanced L2 speakers also made use of semantic properties of the nouns that constituted the collocations, with word sequences containing more concrete nouns receiving lower ratings of association strength.

A separate Bayesian mixed-effects model was fitted for the judgment of association strength, using log Dice as the association measure for the target collocations. Overall, results (see Table 5) replicated the patterns reported when MI was used to measure association strength, which included reliable effects of association strength (log Dice band-low), phrasal frequency band (low), word1 frequency, word1 length, word1 orthographic neighborhood size, word2 length, as well as the interaction between proficiency (L2-advanced) and word2 concreteness. However, given that the effects of medium phrasal frequency band (i.e., phrasal frequency band-medium, L2-intermediate: phrasal frequency band-medium) reported in the MI model were not replicated in the log Dice model, we concluded that L1 and L2 speakers' ratings of association strength might not differ between medium and high phrasal frequency collocations. Both the MI model and the log Dice model found a reliable interaction between proficiency (L2-advanced) and word2 concreteness, indicating that advanced L2 speakers were sensitive to the concreteness of the nouns when judging the association strength of the target collocations. Nevertheless, the effect of word2 concreteness (*M* = 0.047, *SD* = 0.023, 95% CI = [0.000, 0.091]) was only found in the log Dice model. Therefore, the concreteness of the nouns that constituted the collocations may not impact L1 and intermediate L2 speakers' judgment of association strength.

Table 4 Bayesian mixed-effects multinomial modeling results for judgment of association strength (measured by MI)

Estimate	*M*	*SD*	95% CI
MI band-low	-0.638	0.102	[-0.842, -0.437]
Frequency band-low	-0.449	0.081	[-0.609, -0.288]
Frequency band-medium	-0.200	0.064	[-0.321, -0.076]
Proficiency (L2-intermediate): Frequency band-medium	-0.394	0.196	[-0.789, -0.021]
Word1 frequency	-0.266	0.027	[-0.322, -0.216]
Word1 length	-0.042	0.016	[-0.073, -0.012]
Word1 ORTHO	-0.014	0.005	[-0.024, -0.004]
Word2 length	0.040	0.014	[0.014, 0.067]
Proficiency (L2-advanced): Word2 concreteness	-0.089	0.042	[-0.177, -0.008]

Note. Frequency band: band of phrasal frequency. ORTHO: orthographic neighborhood size. Frequency band and MI band were dummy-coded, using the high band as the reference level. Proficiency was dummy-coded, with advanced and intermediate L2 speakers being compared against L1 speakers.

Table 5 Bayesian mixed-effects multinomial modeling results for judgment of association strength (measured by log Dice)

Estimate	*M*	*SD*	95% CI
log Dice band-medium	-0.169	0.059	[-0.291, -0.054]
log Dice band-low	-0.584	0.086	[-0.756, -0.420]
Frequency band-low	-0.235	0.093	[-0.410, -0.042]
Word1 frequency	-0.279	0.025	[-0.329, -0.230]
Word1 length	-0.040	0.015	[-0.070, -0.013]
Word1 ORTHO	-0.022	0.004	[-0.030, -0.013]
Word2 frequency	-0.118	0.028	[-0.171, -0.062]
Word2 length	0.040	0.015	[0.013, 0.069]
Word2 concreteness	0.047	0.023	[0.000, 0.091]
Proficiency (L2-advanced): Word2 concreteness	-0.088	0.041	[-0.170, -0.006]

Note. Frequency band: band of phrasal frequency. ORTHO: orthographic neighborhood size. Frequency band and log Dice band were dummy-coded, using the high band as the reference level. Proficiency was dummy-coded, with advanced and intermediate L2 speakers being compared against L1 speakers.

Discussion

Accuracy of Language Users' Intuitions of Phrasal Frequency and Association Strength

With respect to the accuracy of language users' statistical intuition of collocations, we found that neither L1 nor L2 speakers of English showed accurate intuitions of phrasal frequency and association strength across bands. The inaccuracy of L1 and L2 speakers' intuition of phrasal frequency was consistent with earlier findings. Most previous studies have reported weak or moderate correlations between the subjective judgment of phrasal frequency and objective, corpus-based frequencies of multiword sequences, for both L1 (*r* = .56 in Backman, 1978; *r* = .58 in Siyanova & Schmitt, 2008) and L2 speakers (*r* = .44 in Siyanova & Schmitt, 2008). Splitting phrasal frequency into multiple bands, Siyanova and Schmitt (2008) found that L1 speakers' intuition of phrasal frequency correlated moderately with corpus frequency for medium- (*r* = .742) and high-frequency collocations (*r* = .707). Siyanova-Chanturia and Spina (2015) did not find accurate intuitions of phrasal frequency for medium- and low-frequency collocations among L1 and L2 speakers, yet they concluded both groups of participants' intuitions of high frequency collocations correlated strongly with corpus frequency. We did not find such patterns, and as even for high-frequency collocations, participants' intuitions were still far from accurate (49.9%, 63.3%, and 56.2% for L1 speakers, advanced L2 speakers, intermediate L2 speakers, respectively). Unlike the current study, Siyanova and colleagues (Siyanova & Schmitt, 2008; Siyanova-Chanturia & Spina, 2015) used non-existent collocations as low-frequency materials, either created by L2 learners or researchers. Therefore, it is unclear whether the incorporation of such stimuli might have altered participants' judgment behavior and accuracy. Furthermore, Siyanova-Chanturia and Spina (2015) calculated Cohen's kappa for each collocation as a measure of agreement between corpus-based frequency and subjective frequency estimation, by comparing L1 or L2 speakers' ratings of phrasal frequency against the collocation frequency band. Given that there was no variability in corpus-based frequency band of any given collocation (e.g., "next year" was labeled as high-frequency collocation based on corpus data), we believe Cohen's kappa may not have been an ideal choice for their purpose.

The inaccuracy of language users' intuition of phrasal frequency appears to contradict the findings regarding their intuition of word frequency. As reviewed in the beginning section of this paper, most studies on word frequency intuition (Backman, 1976; Balota et al., 2001; Carroll, 1971; Ringeling, 1984; Shapiro, 1969; Tryk, 1968) indicated that language users—especially L1 speakers—demonstrate accurate intuition of frequency of single words. However, the robustness of language users' intuition of word frequency has also been questioned in recent years (Alderson, 2007; Schmitt & Dunham, 1999), with evidence showing that intuitively estimating how often a word is used in the society is a daunting task even for professional linguists (Alderson, 2007). Such a discrepancy largely results from inconsistencies in methodological choices, as current studies in the literature vary vastly in terms of the choice of stimuli (e.g., the range of frequency and lexical characteristics), corpora, and task (e.g., the magnitude estimation task, the multiple rank order task, or simply classifying words or phrases into frequency bands). Furthermore, evidence supporting or against the robustness of language users' intuition of word frequency have been exclusively built upon correlation coefficients. The choice of correlation coefficients as a measure of agreement or accuracy is problematic, given: 1) cutoff points for labeling correlation coefficients as "weak", "moderate", or "strong" relationship are often arbitrary and inconsistent; 2) correlation coefficients describe the strength and direction of an association between variables, but they do not necessarily reflect the strength of agreement or the degree of accuracy (Schober, Boer, & Schwarte, 2018). Consequently, a strong correlation can still be obtained even when participants' intuitive ratings of phrasal frequency and association strength consistently deviate from corpus-based data. The same issue arises for studies focusing on language users' intuition of phrasal frequency of multiword sequences. Therefore, to validate whether reliable intuition of word or phrasal frequency can be found, more studies that adopt measurements of accuracy that circumvent the shortcomings of correlation coefficients should be carried out. Inconsistencies regarding the accuracy of statistical intuitions also echo a long-lasting debate over whether humans can accurately estimate the statistical information underlying natural events and language use. According to Zacks and Hasher (2002), people automatically track and encode frequencies and probabilities, and their estimation of frequency and probability is accurate, regardless of age, practice, and task manipulations. On the other hand, scholars—especially those in the field of decision-making (Tversky, 1974)—hold that judgment of frequency and probability is unavoidably error-prone, because it involves not only the retrieval of statistical representations, but also task-irrelevant factors, such as the use of strategies (judgmental heuristics). From this point of view, our finding of the inaccuracy of L1 and L2 speakers' intuition of phrasal frequency, as well as reports against the robustness of language users' intuition of word frequency, is not surprising. Despite that no previous work has examined language users' intuition of association strength of multiword sequences, given that L1 and L2 speakers are tuned to both types of statistical information during language use, it is reasonable to assume that results for subjective judgment of association strength should be similar to those for the estimation of phrasal frequency.

Our study also revealed that the accuracy of L1 and L2 speakers' intuition of frequency and association strength followed similar increasing patterns as corpus-based phrasal frequency and association strength increased (except for intermediate L2 speakers' judgment of association strength for medium and high band collocations). Such results partially replicated the findings of Siyanova-Chanturia and Spina (2015) that L1 and L2 speakers showed more accurate intuitions of frequency for highly frequent collocations than for medium and low frequency word combinations. Meanwhile, our results add to the literature in that such increasing pattern may also apply to the whole continuum of both phrasal frequency and association strength. This finding appears to be in line with usage-based accounts and the statistical learning theory (e.g., Gries & Ellis, 2015; Siegelman, 2020), supporting that linguistic knowledge—including statistical intuitions—is acquired from experience: in this case, more frequent or more associated multiword sequences are accompanied with stronger statistical representations. Lastly, differences in the accuracy of statistical intuitions between L1 and L2 speakers are worth mentioning. Siyanova-Chanturia and colleagues (Siyanova & Schmitt, 2008; Siyanova-Chanturia & Spina, 2015) found that L1 speakers demonstrated more accurate intuitions of phrasal frequency than L2 speakers. Moreover, more experienced (advanced) L2 speakers have an advantage compared to less experienced (intermediate) L2 speakers in their judgments of very infrequent word sequences. Our findings are much more complicated than those. With respect to intuitions of phrasal frequency, L1 speakers had an advantage over L2 speakers, but only for low and medium frequency collocations. Similarly, advanced L2 speakers' intuitions of phrasal frequency were more accurate than intermediate L2 speakers, but only for high frequency colocations. When it comes to intuitions of association strength, interestingly, L1 speakers did not have any advantage over L2 speakers, although we did find that advanced L2 speakers showed more accurate intuitions than intermediate L2 speakers for high association strength collocations. Taken together, our results suggest that the development of L1 and L2 speakers' statistical intuitions of multiword sequences might differ and might not strictly follow the pattern reported by Siyanova-Chanturia and colleagues. Needless to say, further studies will be needed to explore this issue.

Linguistic Influences Underlying Intuitions of Phrasal Frequency and Association Strength

Using Bayesian mixed-effects multinomial modeling, this study found that when judging the phrasal frequency and association strength of English adjective-noun collocations, L1 and L2 speakers not only make use of statistical regularities at the phrasal level, but also orthographic, phonological, and semantic characteristics of the words that constitute the word combinations. Interestingly, combining evidence from separate models using MI and log Dice as the measure of association strength, we concluded that L1 and L2 speakers' intuitive judgments of phrasal frequency might not be affected by corpus-based frequency band of the collocations. Instead, they were found to evaluate the degree of association between the constituent words when judging the phrasal frequency of collocations, with collocations in the high association strength band being rated as being used more frequently in English than those in the low association strength band. By contrast, when judging the association strength of collocations, L1 and L2 speakers' subjective estimations were affected not only by corpus-based association strength band, but also by corpus-based frequency band of the collocations. More specifically, collocations in the low association strength band were perceived as being associated less strongly than those in the high association strength band; similarly, low frequency band collocations were also rated as being associated less strongly than those in the high phrasal frequency band. Siyanova-Chanturia and Spina (2015) reported that both L1 and L2 speakers were sensitive to corpus frequency of noun-adjective Italian collocations across bands when instructed to judge how often each word combination was used. However, they did not consider association strength as an additional source of information that could impact language users' intuitive judgment of phrasal frequency. Although more studies will be needed to validate our findings, we interpret such results as evidence indicating that intuitive judgments of phrasal frequency and association strength of multiword sequences might reflect distinct cognitive processes. For adjective-noun collocations, L1 and L2 speakers' intuition of phrasal frequency is driven by their knowledge of the degree of association strength and lexical characteristics of the constituent words (which will be discussed in the following paragraph), whereas L1 and L2 speakers' intuitive judgments of association strength are based on their knowledge of the degree of association strength and phrasal frequency, as well as lexical characteristics of the constituent words. Substantial psycholinguistic evidence (e.g., Arnon & Snider, 2010; Yi, 2018; Yi et al., 2017; Wolter & Gyllstad, 2013) has shown that L1 and L2 speakers are sensitive to frequencies of multiword sequences during online tasks, with more frequent word combinations being processed significantly faster than less frequent ones. Consequently, the absence of effects of corpus-based collocation frequency on L1 and L2 speakers' intuitions of phrasal frequency as reported here might relate to the explicit nature of the forced-choice judgment tasks used in the current study.

Lexical characteristics of words that constitute the collocations were also found to contribute to L1 and L2 speakers' statistical intuitions. In terms of the estimation of phrasal frequency, both L1 and L2 speakers (advanced or intermediate) were found to make use of orthographic (i.e., word1 length, word1/word2 orthographic neighborhood size), phonological (i.e., word2 phonological neighborhood size), and semantic (i.e., word2 concreteness) information of the constituent words in addition to their frequencies (i.e., word1 frequency, word2 frequency). Siyanova-Chanturia and Spina (2015) incorporated word1 frequency and word2 frequency into their analysis, yet neither effect was significant. In the present study, L1 and L2 speakers rated target collocations as being used more frequently if they contained higher-frequency constituent words. The reliable yet negative effect of word1 length replicated that of Siyanova-Chanturia and Spina (2015). Nevertheless, these scholars used noun-adjective Italian collocations, instead of adjective-noun English collocations as in our case. Taken together, such results suggest that it might be the length of the first constituent word—regardless of its part of speech—that impacts language users' estimation of the collocation it constitutes. Specifically, adjective-noun or noun-adjective collocations containing longer first constituent words tend to be perceived as being of lower frequency.

Balota et al. (2001) did not find any effect of orthographic neighborhood size for L1 speakers' subjective estimation of word frequency. However, in the current study, adjective-noun collocations consisting of adjectives with more orthographic neighbors and nouns with more phonological neighbors received higher ratings of phrasal frequency. Regardless of the choice of the measure of association strength (MI or log Dice), for L1 and advanced L2 speakers, collocations containing nouns with more orthographic neighbors and being more concrete would be perceived as being used less frequently. Interestingly, such a pattern was reversed for intermediate L2 speakers. Orthographic and phonological neighborhood size have been found to facilitate visual word recognition (Andrews, 1997; Yates, Locker, & Simpson, 2004). However, it is not clear how these factors impact language users' intuition of statistical regularities. Compared with the intuition of phrasal frequency, L1 and L2 speakers' judgments of association strength were less influenced by lexical characteristics of the constituent words. For both L1 and L2 speakers (advanced or intermediate), effects of word1 frequency, word1 length, and word1 orthographic neighborhood size indicated that target collocations containing adjectives that were less frequent, shorter, and with fewer orthographic neighbors tended to be rated as being of stronger association. By contrast, the effect of word2 length suggested that target collocations with shorter nouns would be perceived as being associated less strongly. Combining the evidence from the two models using MI and log Dice as the measure of association strength, we conclude that semantic concreteness also impacted advanced L2 speakers' intuition of association strength, with target collocations consisting of more concrete nouns receiving lower ratings of association strength. Needless to say, given the exploratory nature of the current study, the intriguing patterns regarding the impact of lexical characteristics on language users' statistical intuitions will need to be validated by future research. Last but not least, the contribution of linguistic characteristics of collocations and their constituent words as reported here also echoes the debate on the holistic vs. analytic processing of multiword sequences (for a review, see Siyanova-Chanturia, 2015). Specifically, the contributions of linguistic characteristics at the single-word and phrase levels to L1 and L2 speakers' intuitive judgment of phrasal frequency and association strength seem to support the analytic processing of constituent words in addition to the holistic processing of word sequences, as revealed by some reaction time studies.

The Relationship between Intuitions of Phrasal Frequency and Association Strength

The present study is the first that simultaneously examined language users' intuitions of phrasal frequency and association strength of multiword sequences. As reported in the previous section, L1 and L2 speakers made use of association strength when asked to judge collocation frequency, and they employed both phrasal frequency and association strength when asked to judge the association strength of collocations. Such results indicate that language users' intuitions of phrasal frequency and association strength may not be separable, thus echoing previous work on the processing of multiword sequences. For instance, Yi and colleagues (Yi, 2018; Yi et al., 2017) found that both phrasal frequency and association strength contribute to L1 and L2 speakers' online processing of multiword sequences. Moreover, given L1 and L2 speakers also accessed the co-occurrence probability (i.e., association strength) of collocations when judging phrasal frequency, we suggest that future studies on statistical intuitions consider association strength as an important, non-negligible variable. Although phrasal frequency and association strength of multiword sequences seem closely related, our study indicates that intuitive estimation of phrasal frequency and association strength might not follow the same cognitive processes. Intuitive judgment of collocation frequency relies on the access to the knowledge of association strength, as well as orthographic, phonological, and semantic characteristics of both constituent words. By contrast, when intuitively estimating the association strength of adjective-noun collocations, L1 and L2 speakers retrieve both phrasal frequency and association strength information, with less reliance on linguistic properties of the constituent words.

Conclusion

In conclusion, this study extended previous work in the literature by investigating L1 and L2 speakers' intuitions of phrasal frequency and association strength of multiword sequences. Our data showed that L1 and L2 speakers' statistical intuitions of collocations are not accurate. Furthermore, their intuitive knowledge of phrasal frequency and association strength seems related, despite following different cognitive processes. With regard to the linguistic influences underlying language users' statistical intuitions, we found that knowledge of both multiword sequences and their constituent words are accessed, with orthographic, phonological, and semantic properties of the constituent words playing important roles, especially for the judgment of phrasal frequency. From a practical point of view, our results do not support the practice of using intuitive estimations as surrogates for corpus-based statistics when selecting multiword sequences for teaching and research purposes. From a theoretical perspective, our finding of the inaccuracy of L1 and L2 speakers' statistical intuitions of collocations does not go against the well-established effects of statistical regularities during the online processing of multiword sequences. Instead, we take such a discrepancy as evidence supporting the existence of two distinct types of statistical knowledge of larger-than-word units: statistical intuitions captured by metalinguistic judgment tasks are explicit and error-prone, whereas statistical representations activated during online processing tasks are implicit and highly automatic.

This study was not without limitations. As pointed out by an anonymous reviewer, almost all target collocations were combinations of nouns with qualifier adjectives (e.g., good, hot, young) with exceptions of a few indefinite adjectives (i.e., other, certain). Such non-homogeneity might have altered the speakers' intuitions because qualifying adjectives and indefinite adjectives differ in the way they describe nouns. Additionally, given that we did not consider individual learner differences other than language proficiency in the present study, future research could be carried out to explore how individual learner differences impact language users' statistical intuitions. It is also worth noting that the choice of corpora could impact the results of research on statistical intuitions at the single word and multiword level. Corpora consist of extensive collections of samples of word usage and are believed to be a good representation of language users' linguistic experience (Balota et al., 2001). Nevertheless, the degree of representativeness may vary enormously depending on the size and design of corpora. Brysbaert and New (2009) concluded that a corpus of 1-3 million words suffices for reliable estimates for high-frequency words (frequency > 20 per million). For low-frequency words (frequency < 10 per million), a corpus of at least 16 million words is needed to allow researchers to get reliable frequency norms. Gablasova et al. (2017) explored the impact of genres (i.e., academic writing, news, fiction), registers (i.e., formal vs. informal), and modality (i.e., written vs. spoken) on the association strength of collocations, using subcorpora of the BNC. They found that collocational strength varied considerably across linguistic settings. Furthermore, they suggested that extra attention should be paid when investigating the link between L2 speakers' linguistic experience and their collocational knowledge, given that their exposure to L2 is likely to be limited and imbalanced across different domains. To advance this field of study, future research should critically evaluate to what extent the corpora are representative of the input L1 and L2 speakers receive based on the above-mentioned dimensions. Lastly, given that studies on intuitions of statistical regularities—especially association strength—of multiword sequences are still lacking, more studies will be needed to validate our research findings.

Notes

1. We referred to information provided by Educational Testing Services on how scores from each section in TOEFL iBT can be interpreted based on an overall proficiency scale of "Below low-intermediate" to "Advanced". To qualify for High-intermediate, one must at least possess a total score of 72 (18 for reading, 17 for reading, 20 for speaking, and 17 for writing); similarly, one is considered to belong to Advanced if one possesses at least a total score of 95 (24 for reading, 22 for listening, 25 for speaking, and 24 for writing). The information was retrived at https://www.ets.org/toefl/test-takers/ibt/scores/understanding/

2. Potential scale reduction factor (PSRF, or R-hat) is a statistical index used to diagnose convergence of MCMC chains. If PSRF is close to 1, we can conclude that MCMC chains are well converged, and the parameter estimates are valid based on the converged chains (Gelman et al., 2013).

3. MI (Yi et al., 2017) and log Dice (Öksüz et al., 2021) scores were computed based on the following mathematical formulas:

$$\mathrm{MI} = \log_2 \frac{f(xy) \times N}{f(x) \times f(y)}$$

$$\text{log Dice} = 14 + \log_2 \frac{2 \times f(xy)}{f(x) + f(y)}$$

N is the size of the corpus. f(xy), f(x) and f(y) refer to the frequency of the collocation/the node (i.e., the adjectives)/the collocate (i.e., the nouns) in the whole corpus, respectively.

References

Alderson, J. C. (2007). Judging the frequency of English words. *Applied Linguistics, 28*, 383–409. https://doi.org/10.1093/applin/amm024

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review, 4*, 439–461. https://doi.org/10.3758/BF03214334

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

Backman, J. (1976). Some common word attributes and their relations to objective frequency counts. *Scandinavian Journal of Educational Research*, *20*, 175–186. https://doi.org/10.1080/0031383760200112

Backman, J. (1978). Subjective structures in linguistic recurrence. ERIC Document ED180195.

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, *29*, 639–647. https://doi.org/10.3758/BF03200465

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911. https://doi.org/10.3758/s13428-013-0403-5

Carroll, J. B. (1971). Measurement properties of subjective magnitude estimates of word frequency. *Journal of Verbal Learning and Verbal Behavior*, *10*, 722–729. https://doi.org/10.1016/S0022-5371(71)80081-6

Carroll, J. B., Davies, P., & Richman, B. (1971). *Word frequency book*. Boston, MA: Houghton Mifflin.

Diependaele, K., Lemhofer, K., & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology*, *66*, 843–863. https://doi.org/10.1080/17470218.2012.720994

Dimroth, C. (2018). Input and the acquisition of productive grammatical knowledge: Vocabulary size as missing link? *Linguistic Approaches to Bilingualism*, *8*, 717–721. https://doi.org/10.1075/lab.18057.dim

Ellis, N. C., & Ogden, D. C. (2017). Thinking about multiword constructions: Usage-based approaches to acquisition and processing. *Topics in Cognitive Science*, *9*, 604–620. https://doi.org/10.1111/tops.12256

Ellis, N. C., Romer, U., & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar* (Language Learning Monograph Series). Malden, MA: Wiley.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text & Talk*, *20*, 29–62. https://doi.org/10.1515/text.1.2000.20.1.29

Fletcher, W. H. (2011). Phrases in English (PIE). Retrieved from http://phrasesinenglish.org/

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Boca Raton, FL: CRC Press.

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

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, *33*, 1–22. http://dx.doi.org/10.18637/jss.v033.i02

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge, UK: Cambridge University Press.

Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *Plos One, 7*. https://dx.doi.org/10.1371%2Fjournal.pone.0043230

McCrostie, J. (2007). Investigating the accuracy of teachers' word frequency intuitions. *Regional Language Centre Journal*, *38*, 53–66. https://doi.org/10.1177%2F0033688206076158

McDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, *43*, 1735–1751. https://doi.org/10.1016/S0042-6989(03)00237-2

Öksüz, D., Brezina, V., & Rebuschat, P. (2021). Collocational processing in L1 and L2: The effects of word frequency, collocational frequency, and association. *Language Learning*, *71*, 55–98. https://doi.org/10.1111/lang.12427

Plummer, M., Best, N., Cowles, K., & Vines, K. (2005). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, *6*, 7–11.

R Core Team. (2021). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Ringeling, T. (1984). Subjective estimations as a useful alternative to word frequency counts. *Interlanguage Studies Bulletin*, *8*, 59–69.

Schmitt, N., & Dunham, B. (1999). Exploring native and non-native intuitions of word frequency. *Second Language Research*, *15*, 389–411. https://doi.org/10.1191%2F026765899669633186

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia and Analgesia*, *126*, 1763–1768. https://doi.org/10.1213/ane.0000000000002864

Schwanenflugel, P. (1991). Why are abstract concepts hard to understand? In P. J. Schwanenflugel (Ed.), *The psychology of word meaning* (pp. 223–250). Mahwah, NJ: Erlbaum.

Shapiro, B. J. (1969). Subjective estimation of relative word frequency. *Journal of Verbal Learning and Verbal Behavior*, *8*, 248–251. https://doi.org/10.1016/S0022-5371(69)80070-8

Siegelman, N. (2020). Statistical learning abilities and their relation to language. *Language and Linguistics Compass*, *14*(3), e12365. https://doi.org/10.1111/lnc3.12365

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

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

Siyanova-Chanturia, A., & Spina, S. (2015). Investigation of native speaker and second language learner intuition of collocation frequency. *Language Learning*, *65*, 533–562. https://doi.org/10.1111/lang.12125

Thorndike, E. L., & Lorge, I. (1944). *A teacher's word book of 30,000 words*. New York, NY: Columbia University Press.

Tryk, H. E. (1968). Subjective scaling of word frequency. *American Journal of Psychology*, *81*, 170–177. https://psycnet.apa.org/doi/10.2307/1421261

Tversky, A. K. D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131. https://psycnet.apa.org/doi/10.1126/science.185.4157.1124

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. *Studies in Second Language Acquisition*, *35*, 451–482. https://doi.org/10.1017/S0272263113000107

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*, 395-416. https://doi.org/10.1017/S0272263117000237

Yates, M., Locker, L., & Simpson, G. B. (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin & Review*, *11*, 452–457. https://doi.org/10.3758/BF03196594

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

Yi, W., Lu, S., & Ma, G. (2017). Frequency, contingency and online processing of multiword sequences: An eye-tracking study. *Second Language Research, 33*(4), 519-549. https://doi.org/10.1177/0267658317708009

Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *Frequency processing and cognition* (pp. 21–36). Oxford, UK: Oxford University Press.

Supporting Information

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

Appendix S1. Summary of studies on language users' intuition of word frequency

Appendix S2. Adjective-noun collocations used in the current study

Appendix S3. L2 participants' demographic information

Appendix S4. Characteristics of the selected collocations

Appendix S5. Correlations among characteristics of the collocations

Appendix S6. Questionnaire of intuitions of phrasal frequency and association strength

Appendix S7. Accuracy of L1 and L2 speakers' statistical intuitions across bands

Appendix S8. Accuracy of participants' intuition of phrasal frequency

Appendix S9. Accuracy of participants' intuition of association strength (measured by MI)

Appendix S10. Accuracy of participants' intuition of association strength (measured by log Dice)