LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Investigating First and Second Language Speaker Intuitions of Phrasal Frequency and Association Strength of Multiword Sequences

Wei Yi [a], Kaiwen Man, [b] and Ryo Maie [c]

[a]Peking University [b]University of Alabama [c]Michigan State University

Abstract: In this study, we investigated the accuracy of first language (L1) and second language (L2) speakers' intuitive judgments of phrasal frequency and collocation strength, and examined the linguistic influences that give rise to these judgments. L1 and L2 speakers of English judged 180 adjective–noun collocations as (a) high frequency, medium frequency, or low frequency and (b) high association, medium association, or low association. Results showed that neither L1 nor L2 speakers demonstrated accurate intuitive judgments of phrasal frequency and association strength. Both groups of participants employed linguistic information at phrase and single-word levels when giving intuitive statistical estimates. We found judgments of phrasal frequency and association strength to be intertwined for both L1 and L2 speakers. Taken together, these findings shed new insight on understanding language users' statistical knowledge of multiword sequences.

CRediT author statement – **Wei Yi**: conceptualization; methodology; investigation; writing – original draft preparation; writing – review & editing. **Kaiwen Man**: conceptualization; formal analysis; writing – review & editing. **Ryo Maie**: writing – original draft preparation (literature review); writing – review & editing.

A one-page Accessible Summary of this article in non-technical language is freely available in the Supporting Information online and at https://oasis-database.org

We would like to express our gratitude to Professor Michael Long for his support for this project. We also gratefully acknowledge the insightful and constructive feedback from Journal Editor Professor Emma Marsden and Associate Editors Professor Scott Crossley and Professor Kristopher Kyle, as well as the anonymous 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. Email: weiyisla@pku.edu.cn

The handling editor for this manuscript was Kristopher Kyle.

DOI: 10.1111/lang.12521

Keywords phrasal frequency; association strength; collocations; multiword expressions; statistical intuition

Introduction

Judgments of the frequency and the probability of events have survival benefits for humans in daily activities. As part of human cognition, the use of language makes no exception in this regard. Natural languages abound 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 first language (L1) and second language (L2) speakers demonstrate reliable statistical intuitions. However, researchers have primarily focused their efforts on the accuracy of intuitive judgments of word frequency, leaving unclear whether language users possess accurate intuition for 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 in our study, we explored L1 and L2 speakers' intuitions of two kinds of statistical information about 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 constitute such larger-than-word units along with corpus statistics at the phrase level.

Background Literature

Intuition of Frequency of Single Words

Frequency of occurrence is a fundamental piece of information that people encode about their experience with language. Frequency of occurrence indicates how likely language users are to encounter a linguistic unit and determines the degree of automaticity with which language users process or retrieve a word or phrase (Gries & Ellis, 2015). Researchers have studied frequency effects extensively from various perspectives. For lexical processing, ample evidence has shown that L1 and L2 speakers are sensitive to the frequency of single words (Diependaele et al., 2013). Along with research on frequency, in many studies, researchers have also examined, from practical and theoretical concerns, language users' intuitive estimation of word frequency. 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);

applied linguists need 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 L2 acquisition, many researchers (e.g., Ellis & Gries, 2015; Ellis, Romer, & O'Donnell, 2016) have held that language users are attuned to frequency of input and that language users can acquire knowledge of frequency and other probabilistic information after decades of language use.

Research on language users' intuition about word frequency was initiated in studies conducted in the 1960s and 1970s (for a summary of studies on intuitive judgments of word frequency, see Appendix S1 in the Supporting Information online). These studies typically followed a three-stage paradigm. First, researchers extracted objective frequency counts of preselected lexical items from corpora that were assumed to reflect the language use of the society from which they were drawn. Subsequently, researchers asked participants to judge how often the words were used in a language community or in their personal experience, collecting participants' responses through the magnitude estimation method or through a multiple rank order task (Shapiro, 1969). A magnitude estimation task requires participants to estimate how frequently words occur by assigning numbers to each word based on certain response scales; a multiple rank order task asks participants to rank words according to the relative frequency of the words. Finally, to evaluate the accuracy of language users' intuition of word frequency, researchers calculated Pearson product-moment correlations for magnitude estimations or Spearman rank-order correlations for multiple rank order tasks between these subjective frequency estimates and objective frequency counts from corpora.

Tryk (1968) sampled 100 nouns to represent the spectrum of the Thorndike-Lorge frequency counts and asked undergraduate students to estimate how often average English speakers used each noun in daily conversation in a given period of time. Correlations between subjective frequency estimates and log-transformed objective frequency counts were moderately high, ranging from .74 to .78. Shapiro (1969) selected 91 words from the Thorndike-Lorge (Thorndike & Lorge, 1944) and Kučera-Francis (Kučera & Francis, 1967) tabulations and had six groups of L1 English speakers estimate the frequencies of the words in either spoken or written language. Participants first ranked the words based on their intuition. Shapiro (1969) then asked the participants to assign numbers to all the words in the entire list according to the relative frequency of the words. 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's (1969) list and found that L1 speakers' estimation of the frequency of the words correlated highly with objective counts. Backman (1976) took 50 words translated from Shapiro's (1969) list and had L1 speakers of Swedish rate these words for frequency of occurrence in written language. Similarly, Backman found subjective frequency ratings for these 50 words to be highly correlated with objective frequency counts of the words (.93).

In later studies, researchers have reported L1 speakers' accurate intuition of word frequencies. Ringeling (1984) asked 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 et al. (2001) collected subjective frequency estimates for 2,938 monosyllabic English words from L1 speakers. Participants rated each word for how frequently they had encountered it in general and in different domains. In line with previous studies, Balota et al. found relatively high correlations between subjective frequency estimates and objective corpora frequencies (.78–.83). McCrostie (2007) compared the intuition of word frequency of English instructors to that of undergraduate students. McCrostie asked both groups of participants to arrange two lists of words in frequency order. Intriguingly, English teaching professionals' frequency judgments were no better than those of undergraduate students. Moreover, for both groups, their mean judgment accuracy of words in the middle frequency range (.49–.51) was much lower than for words in the whole frequency range (.83–.84).

Unlike the above-mentioned researchers, in their studies, Alderson (2007) and Schmitt and 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 asked L1 and L2 speakers of English to assign frequency ratings relative to an anchor word within each lexical set. In contrast to findings in previous studies, they found only moderate correlations between subjective ratings of word frequency and objective word frequencies from corpus data for L1 speakers (.53) and for L2 speakers (.58). Alderson (2007) conducted three experiments to investigate the accuracy of judgments of word frequency by professional linguists. For the first experiment, Alderson asked participants to report how frequently 100 English verbs occurred in every million words; for the second experiment, Alderson asked participants to

rank order 50 verbs according to their frequency; and for the third experiment, Alderson asked participants to rank order 25 verbs that he had chosen to represent a more evenly distributed range of word frequency. Similar to Schmitt and Dunham (1999), Alderson found only moderate correlations 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 from single words to multiword sequences. Language users widely employ multiword sequences (Erman & Warren, 2000), and multiword sequences play a critical role in achieving nativelike 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 language users encounter a word combination, whereas association strength is a measure of the co-occurrence probability of words that constitute the word combination; association strength is thought to be linked to language users' ability to predict the words following or preceding another word in a sequence (Gablasova et al., 2017). Recent empirical studies have supported the conceptual distinction between phrasal frequency and association strength by showing that phrasal frequency alone is not adequate for explaining 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 that is independent of phrasal frequency (e.g., Yi, 2018; Yi et al., 2017). Phrasal frequency and association strength do not necessarily go hand in hand. In fact, one can even expect to find 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 in the Supporting Information online).

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 et al., 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 MI and log Dice use a logarithmic scale and highlight the exclusivity between words in collocations. MI and log Dice are fairly similar. However, they differ in a number of aspects: (a) MI expresses the ratio between the frequency of collocations and the frequency of random co-occurrences of the constituent words, whereas log Dice captures the tendency of two words to co-occur relative to the frequencies of these words in a corpus (Gablasova et al., 2017); (b) MI does not have a theoretical minimum and maximum, whereas log Dice has a maximum value of 14; (c) 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 (BNC), 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 language users' 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 that have been reported for single words may extend to word combinations as well. So far, few researchers 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 18 three-word combinations. In an approach similar to that used in Schmitt and Dunham's (1999) study, Backman used a word sequence as an anchor and asked 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 for 31 frequent (nativelike) and 31 infrequent (learner) adjective–noun pairings extracted from a learner corpus of writings. Siyanova and Schmitt further divided the frequent collocations into high- and medium-frequency bands based on a cutoff point of frequency at 100 occurrences in the BNC. They asked participants to rate each collocation on a scale of 1 (*very uncommon*) to 6 (*very common*). Overall, participants' ratings of phrasal frequency did not correlate highly with corpus-based data for L1 speakers (.58) or for L2 speakers (.44). Interestingly, Siyanova and Schmitt found that L1 speakers were able to distinguish frequent collocations from infrequent collocations or 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 created another group of noun–adjective combinations that were incorporated as low-frequency collocations. Siyanova-Chanturia and Spina instructed L1 and L2 speakers of Italian to report their intuition of phrasal frequency based on a 4-point frequency scale (i.e., *high*, *medium*, *low*, *very low*). Results showed that frequency of collocations in the corpus predicted both L1 and L2 speakers' judgments of collocation frequency. Using Cohen's kappa to measure the agreement between language users' judgments of phrasal frequency and objective corpus-based frequency bands for each item, Siyanova-Chanturia and Spina 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 processing multiword sequences, recent evidence has suggested that language users may access the knowledge of word combinations as well as the constituent parts of the combinations (for a review, see Siyanova-Chanturia, 2015). For instance, researchers found L1 and L2 speakers 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 about words. If knowledge of constituent words contributes to the processing of multiword sequences, one might 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 frequencies correlated with word pronounceability, defined as the degree of difficulty in pronouncing a word (.82), and with comprehensibility, defined as the degree of difficulty in comprehending a word (.65). 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. It is interesting that neighborhood size also contributed to subjective frequency ratings, but only for highly familiar items. Siyanova-Chanturia and Spina (2015) have been the only researchers who have 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.

The Present Study

The current literature is limited in the following aspects. First, although there have been ample studies in which researchers examined the accuracy of language users' intuition of frequency of occurrence, most studies have focused on single words, leaving unclear whether similar patterns could be found for larger-than-word units. Second, in no single study have researchers explored language users' intuition of association strength of multiword sequences. Yet, research has shown that language users' knowledge of association strength and phrasal frequency are related (Yi, 2018; Yi et al., 2017). As Siyanova-Chanturia and Spina (2015) acknowledged, 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 on which language users rely when making subjective judgments about frequency and probability of language use; research is especially lacking for how knowledge of constituent words contributes to statistical intuitions of multiword sequences. Fourth, studies in which researchers have examined L2 speakers' statistical intuition have been scarce.

To bridge these gaps, we explored in our study 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. We specifically chose MI and log Dice as measures of association strength of collocations. As we mentioned previously, MI and log Dice capture the bidirectional relationship between constituent words in collocations and operate on

normalized scales, which could make our results comparable across corpora. Moreover, the choice of log Dice allowed us to examine the exclusivity of collocations without the low-frequency bias that occurs in MI (Gablasova et al., 2017). In addition to corpus-retrieved word frequency, phrasal frequency, and association strength, we included word length (i.e., 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. We chose word length (Siyanova-Chanturia & Spina, 2015) and neighborhood size (Balota et al., 2001) because research has shown that they moderate intuitions of frequency of words or multiword sequences, and we selected concreteness 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 judgments 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?

Method

Participants

We recruited 194 participants, including 81 L2 English learners (55 females) and 113 L1 English speakers (58 females). The L2 speakers were Chinese international students studying in undergraduate programs in the United States. The L1 speakers were residents in the United States and had earned at least a bachelor's degree at the time of data collection. The mean ages of the L1 and L2 speakers were 34.1 years ($SD = 12.7$) and 24.0 years ($SD = 3.5$), respectively. The L2 speakers' mean age of onset for learning English was 9.0 years ($SD = 2.7$), and their mean length of residence in the United States 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, we classified the L2 participants following the TOEFL official guide[1] as either 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. Based on 5-point scales, L2 speakers' mean self-reported English use outside the classroom was 3.4 ($SD = 0.8$), and their mean ratings of English proficiency were 3.3 ($SD = 0.8$) for reading, 3.2 ($SD = 0.8$) for listening, 2.9 ($SD = 0.8$) for speaking, and 2.9 ($SD = 0.8$) for writing (see Appendix S3 in the Supporting Information online for more information about the L2 participants' characteristics).

Stimuli

We borrowed 180 English adjective–noun collocations from those used in Yi's (2018) study that we retrieved from Phrases in English (Fletcher, 2011), an online database. Phrases in English was derived from the second edition of the BNC, which is a balanced corpus consisting of 100 million words of modern British English that are widely present in the written and spoken domains. Following Wolter and Gyllstad (2013), we defined collocations as multiword sequences consisting of words that co-occur more frequently than chance would predict given the frequency of the constituent words. Yi (2018) specifically defined adjective–noun combinations as collocations if (a) they occurred at least once per million words in the BNC and if (b) the statistical association between an adjective and a noun measured by MI was higher than 3.0.

We sampled the collocations from the BNC, via Phrases in English, such that they represented the whole range of frequency and association strength (i.e., MI) of adjective–noun combinations that met the above-mentioned criteria. On the basis of frequencies retrieved from the BNC, we computed MI and log Dice values for each collocation.[2] To ensure that the target collocations were familiar to our L2 participants, we asked five intermediate-to-advanced L2 speakers of English who did not participate in this study to rate their familiarity with the collocations on a scale of 1 (*totally unknown*) to 5 (*extremely familiar*). The mean familiarity rating was 4.5 ($SD = 0.4$) for our target collocations. To address the linguistic influences responsible for language users' intuitive judgments of phrasal frequency and association strength, we retrieved lexical properties of the constituent words (i.e., variables word1 and word2), including word length, orthographic neighborhood size, and phonological neighborhood size, from the CLEARPOND database (Marian et al., 2012), which provides an interface for obtaining phonological and orthographic neighborhood sizes across languages. Furthermore, we borrowed concreteness ratings of the nouns within each collocation from Brysbaert et al.'s (2014) study.

Language usage can differ between British and American English. To evaluate differences in usage, we looked up phrasal frequency and association strength of collocations, as well as frequencies of the constituent words, in

the Corpus of Contemporary American English (Davies, 2008), which is a balanced, large-scale corpus consisting of around one billion words. Overall, corpus data obtained from the BNC and the Corpus of Contemporary American English were highly correlated (.73 for phrasal frequency, .78 for MI, .68 for log Dice, .90 for the frequency of the adjectives, and .87 for the frequency of the nouns). We transformed frequencies to occurrences per million words before transforming the result to their natural logarithm. Based on the BNC data, we then ranked the collocations from the lowest to the highest phrasal frequency and grouped them into three separate lists of high, medium, or low phrasal frequency collocations; each list contained 60 items. We adopted the same procedure for association strength. This resulted in three bands of collocations categorized into high, medium, or low association strength. For grouping collocations for phrasal frequency and association strength, no borderline items existed (for a full list of the collocations, see Yi et al., 2022c, and Appendix S2 in the Supporting Information online). The mean phrasal frequencies of collocations were 0.331 ($SD = 0.173$, *range* = 0.077–0.663) in the low band, 1.284 ($SD = 0.159$, *range* = 1.072–1.896) in the medium band, and 2.426 ($SD = 0.420$, *range* = 1.984–3.780) in the high band. On average, the association strengths of collocations, measured by MI, were 5.108 ($SD = 0.834$, *range* = 3.366–6.157) in the low band, 7.201 ($SD = 0.601$, *range* = 6.179–8.101) in the medium band, and 9.511 ($SD = 1.199$, *range* = 8.180–12.713) in the high band. The mean association strengths of collocations, measured by log Dice, were 6.855 ($SD = 0.711$, *range* = 4.763–7.761) in the low band, 8.362 ($SD = 0.316$, *range* = 7.803–9.009) in the medium band, and 9.815 ($SD = 0.648$, *range* = 9.010–11.495) in the high band. Appendix S4 in the Supporting Information online provides characteristics of the selected collocations, and Appendix S5 shows the correlations among characteristics of the collocations.

Instrument and Task

We incorporated the collocations into a questionnaire with three sections. In the first section, the participants answered several questions about their demographic information. In the second section, the participants judged how frequently a collocation is used in English based on a 3-point scale (*low frequency*, *medium frequency*, and *high frequency*). In the final section, the participants estimated how strongly two words are associated within a collocation. The strength of association was explained to the participants to be how likely the constituent words can predict the appearance of one word given the other word, regardless of the direction of prediction. The participants responded 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 a multiple rank-order task or the magnitude estimation task, yet it would have been rather unnatural to require the 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), the 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, we randomly grouped the collocations into nine blocks, each containing 20 collocations. We randomized the order of presentation of the blocks of target items to the participants as well as the order of items within each block. To help the participants understand the tasks, we provided them with examples (for the complete questionnaire, see Yi et al., 2022b, and Appendix S6 in the Supporting Information online).

Procedure

We administered the questionnaire online through Qualtrics and Amazon Mechanical Turk. We used Qualtrics to deliver the questionnaire to international Chinese students studying at the undergraduate level in the United States. Amazon Mechanical Turk was a better method for reaching L1 speakers of English considering the large potential pools of respondents. Moreover, Amazon Mechanical Turk allowed us to restrict the participants to L1 English speakers living in the United States and holding a bachelor's degree. We created two versions of the questionnaire to counterbalance the order of the judgment of phrasal frequency and association strength of collocations, and we administered each version of the questionnaire to one half of the participants. The participants were informed that there was no time pressure; they were to make judgments relying on their own intuition, and there was no right or wrong answer. The questionnaire took about 25 minutes for the participants to complete.

Data Analysis

In this study, we followed the suggestion of Alderson (2007) 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, medium, or high phrasal frequency, or low, medium, or high association strength). To reveal how the participants came to their subjective estimations, we ran Bayesian mixed-effects multinomial models separately for phrasal frequency and for association strength measured by MI and log Dice. We deemed Bayesian mixed-effects multinomial analysis to be an optimal approach because (a) it does not assume that all parameters are normally distributed, (b) it can handle small sample sizes, and (c) it allows users to model outcome variables consisting of three categories. We excluded from data analyses the seven L2 participants who did not report their TOEFL iBT scores. For both sets of statistical models, we included the following predictors: proficiency (L1 speakers, advanced L2 speakers, intermediate L2 speakers), phrasal frequency band (low, medium, high), MI band (low, medium, high), log Dice band (low, medium, 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, we added interactions of proficiency with the other variables. We dummy-coded proficiency to compare the intermediate and the advanced L2 speakers to the L1 speakers. Similarly, we dummy-coded phrasal frequency, that is, MI band and log Dice band, using the high band as the reference level. Our incorporation of MI band and log Dice band as predictors when we modeled the participants' subjective judgments of phrasal frequency enabled us to explore whether the participants made use of statistical association information when they judged the phrasal frequency of collocations. We applied the same rationale to the inclusion of phrasal frequency band when we modeled the participants' subjective judgments of association strength. Given that we had already included orthographic, semantic, and phonological variables specific to each item for statistical analyses, we considered only random intercepts of participants for both sets of models.

We used Bayesian estimation for modeling parameter estimation via the MCMCglmm package (Hadfield, 2010) in the R software (R Core Team, 2021). We assessed convergence via potential scale reduction parameters[3] (Gelman et al., 2013) with the coda package (Plummer et al., 2005) for R. We performed two chains using 13,800 iterations with thinning of 4 to reduce autocorrelation among samplers. We summarized model parameter estimates, standard deviations, and their 95% credible intervals based on the posterior densities using the final 10,300 iterations after burn-in 3,500 (Gelman et al., 2013). For our

study, we used a potential scale reduction factor of 1.2 or less for each model parameter as the cut-off indicating convergence (Gelman et al., 2013). We used default noninformative priors offered by the MCMCglmm package (Hadfield, 2010). Specifically, for the fixed effects, we specified flat normal priors with means of zeros and large variances of 10^8. In addition, we used diffuse priors 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). We summarized the Bayesian model results 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 included zero, then it was reasonable to infer that the parameter (variable) very likely could take a value near zero, indicating that it did 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 did not contain zero. Full model results as well as trace plots for model parameters are available via Open Science Framework at https://osf.io/r9avk (see also Yi et al., 2022a, for the datasets used in this study).

Results

Accuracy of L1 and L2 Speakers' Statistical Intuitions

Overall, the results suggested that, for both the L1 and the L2 speakers of English, their subjective intuitions of phrasal frequency and association strength were not accurate. As Figure 1 shows, the participants' judgment accuracy for corpus-based phrasal frequency or for association strength followed an increasing pattern whether measured by MI or log Dice, except for the intermediate L2 speakers' judgment of association strength for medium and high MI band collocations or medium and high log Dice band collocations. The L1 speakers' intuition of phrasal frequency was more accurate than that of L2 speakers, but only for low- and medium-band collocations. However, for intuitions of association strength, the L1 speakers did not have such an advantage over the L2 speakers. Interestingly, the advanced L2 speakers exhibited more accurate intuitions of high-frequency or high-association-strength collocations than did the intermediate L2 speakers, yet such a pattern was reversed for low-association-strength collocations (for accuracies across the bands, see Appendix S7 in the Supporting Information online).

For each of the 180 collocations, we also calculated the proportion of the participants whose ratings of phrasal frequency and of association strength matched the groupings of the corpus-based bands (see Appendices S8, S9,

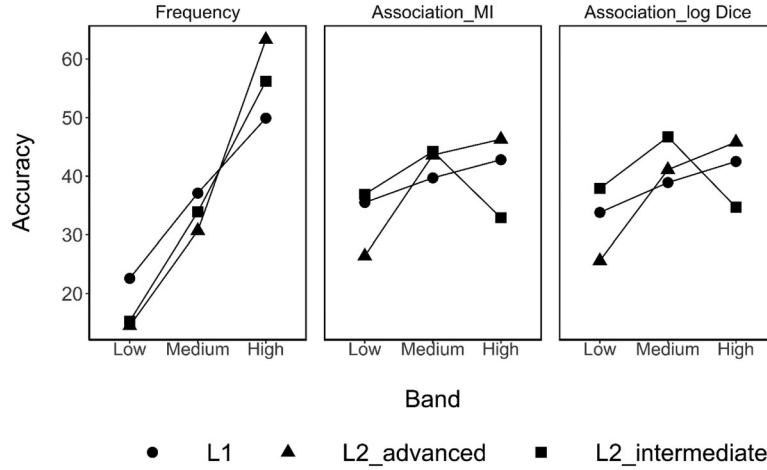

Figure 1 Percent accuracy of L1 and L2 speakers' intuitions of phrasal frequency and association strength measured by mutual information (MI) and log Dice.

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

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

	Proportion of matched responses for individual collocations		
Statistical intuition	[0%, 40%)[a]	[40%, 60%)[a]	[60%, 100%][a]
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. [a] A parenthesis indicates that the point or value was not included in the interval; a bracket indicates that the value was included in the interval. 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, medium, high) fell into each category. MI = mutual information.

Bayesian Model Results for the Judgment of Phrasal Frequency

With 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 of 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 of proficiency and the following variables: word1 length, word1 phonological neighborhood size, word2 orthographic neighborhood size, and word2 concreteness. As summarized in Table 2, the effect of phrasal frequency band indicated that both the L1 and the 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, along with the interactions of

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

Estimate	*M*	*SD*	95% CrI
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 orthography	0.017	0.005	[0.009, 0.026]
Word2 frequency	0.108	0.033	[0.041, 0.174]
Word2 orthography	−0.032	0.006	[−0.043, −0.021]
Word2 phonology	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 phonology	0.007	0.004	[0.000, 0.016]
Proficiency (L2-Intermediate) × Word2 orthography	0.049	0.023	[0.005, 0.094]
Proficiency (L2-Intermediate) × Word2 Concreteness	0.285	0.086	[0.117, 0.446]

Note. 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. CrI = credible interval; frequency band = band of phrasal frequency; orthography = orthographic neighborhood size; phonology = phonological neighborhood size.

proficiency and MI bands, showed that the L1 and the L2 speakers made use of MI band information when they judged the phrasal frequency of collocations. Both the L1 and the advanced L2 speakers assigned lower ratings of phrasal frequency to low-MI-band collocations than to collocations labeled as high-MI-band sequences. However, for the intermediate L2 speakers, we found reversed patterns—compared with high-MI-band word sequences, collocations in the low and medium MI bands received higher ratings of phrasal frequency.

The effects of the word1 and the word2 variables suggested that orthographic, phonological, and semantic properties of words that constituted the

collocations affected both the L1 and the L2 speakers' subjective estimations of phrasal frequency. Specifically, Table 2 shows that the participants rated target collocations containing higher frequency adjectives (word1 frequency) and nouns (word2 frequency) as being more frequently used than target collocations with lower frequency constituent words. The effect of word1 length, along with the interaction of word1 length and proficiency, indicated that the L1 and the 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 the advanced L2 speakers was significantly weaker than that among the L1 and the intermediate L2 speakers.

For orthographic and phonological influences, we found complex patterns. Table 2 shows that the L1 and the L2 speakers judged target collocations as more frequent if they contained adjectives with more orthographic neighbors (word1 orthographic neighborhood size) or if they contained nouns with more phonological neighbors (word2 phonological neighborhood size). However, for word sequences consisting of nouns with more orthographic neighbors, the L1 and the advanced L2 speakers tended to judge these sequences as being used less frequently (word2 orthographic neighborhood size), whereas the intermediate L2 speakers judged them as being used more frequently. Additionally, the advanced L2 speakers also tended to rate target collocations containing adjectives with more phonological neighbors as more frequent (L2-Advanced × Word1 Phonological Neighborhood Size). For the role of semantic characteristics, the effect of word2 concreteness, along with the interaction of proficiency and word2 concreteness, suggested that the L1 and the advanced L2 speakers perceived target collocations consisting of more concrete nouns as being less frequent, yet this pattern was reversed for the intermediate L2 speakers.

We fit a separate Bayesian mixed-effects model 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 that we reported when we used MI to measure association strength. The model 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 of proficiency and word1 length, word2 orthographic neighborhood size, and word2 concreteness. Nevertheless, some effects reported in the MI model (see Table 2) did not appear in this model; similarly, this model also revealed some effects that did not appear in the MI model. Such inconsistent patterns may have resulted from the fact that MI and log Dice capture different aspects of association strength (as

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% CrI
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 orthography	0.011	0.005	[0.001, 0.020]
Word2 frequency	0.091	0.031	[0.032, 0.153]
Word2 orthography	−0.030	0.005	[−0.039, −0.019]
Word2 phonology	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 orthography	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 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. CrI = credible interval; frequency band = band of phrasal frequency; orthography = orthographic neighborhood size; phonology = phonological neighborhood size.

we mentioned previously) and suggested that the choice of association measures did have an impact on the results. Thus, from the effects of phrasal frequency bands in the MI model (i.e., phrasal frequency band-low) and in the log Dice model (i.e., phrasal frequency band-medium), we could not conclude that the L1 and the L2 speakers were sensitive to corpus-based phrasal frequencies when intuitively judging the phrasal frequency of collocations. Similarly, Table 3 shows that we could not confidently conclude that the advanced L2 speakers rated phrasal frequency lower than did the L1 speakers given that we found the effect of speaker only in the log Dice model. Last, the inconsistencies regarding the interactions of proficiency (L2-advanced) and word1

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

Estimate	*M*	*SD*	95% CrI
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 orthography	−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 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. CrI = credible interval; frequency band = band of phrasal frequency; orthography = orthographic neighborhood size.

phonological neighborhood size, word1 frequency, and word2 concreteness suggested that (a) advanced L2 speakers might not necessarily be sensitive to the phonological neighborhood information of the adjectives that constituted the target collocations, (b) the advanced L2 speakers might not differ from the L1 speakers and the intermediate L2 speakers for the degree of sensitivity to the frequency of the adjectives that constituted the target collocations, and (c) the advanced L2 speakers might not differ from the L1 speakers for the degree of sensitivity to the concreteness of the nouns that constituted the target collocations.

Bayesian Model Results for the Judgment of Association Strength

With 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 of proficiency and phrasal frequency band, word1 frequency, word1 length, word1 orthographic neighborhood size, word2 length, and the interaction of proficiency and word2 concreteness. As summarized in Table 4, the effect of MI band indicated that the L1 speakers and the 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 of proficiency (L2-intermediate) and phrasal frequency band (medium), suggested that the L1 and the L2 speakers also made use of corpus-based phrasal frequencies when judging the association strength of word combinations, with the intermediate L2 speakers being more sensitive to such statistical information than the L1 and the advanced L2 speakers. Specifically, collocations in the low and medium phrasal frequency bands received lower ratings of association strength than did those in the high phrasal frequency band. For the influences of lexical characteristics of constituent words, as shown in Table 4, the effect of word1 frequency, of word1 length, and of word1 orthographic neighborhood size suggested that the participants tended to perceive the words of target collocations containing adjectives that were more frequent, longer, and with more orthographic neighbors as being associated less strongly. In addition, the effect of word2 length indicated that the participants rated the words of target collocations containing longer nouns as being associated more strongly than the words of target collocations with shorter nouns. Last, the interaction of proficiency (L2-advanced) and word2 concreteness showed that the 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.

We fit a separate Bayesian mixed-effects model for the judgment of association strength, using log Dice as the association measure for the target collocations. Overall, the results (see Table 5) replicated the patterns reported when we used MI to measure association strength that 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 of proficiency (L2-advanced) and word2 concreteness. However, given that the effects of medium phrasal frequency band (i.e., L2-Intermediate $\times$ Phrasal Frequency Band-Medium) reported in the MI model were not replicated in the log Dice model, we concluded that the L1 and the L2 speakers' ratings of association strength did not differ between medium and high phrasal frequency collocations. Both the MI model and the log Dice model indicated a reliable interaction of proficiency (L2-advanced) and word2 concreteness, indicating that the advanced L2 speakers were sensitive to the concreteness of the nouns when judging the association strength of the target collocations. Nevertheless, as shown in Table 5, we found the effect of word2 concreteness only in the log Dice model. Therefore, the concreteness of the nouns in the collocations might not have impacted the L1 and the intermediate L2 speakers' judgment of association strength.

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

Estimate	*M*	*SD*	95% CrI
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 orthography	−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 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. CrI = credible interval; frequency band = band of phrasal frequency; orthography = orthographic neighborhood size.

Discussion

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

For the accuracy of language users' statistical intuition of collocations, we found that neither the L1 nor the L2 speakers of English showed accurate intuitions of phrasal frequency and association strength across bands. The inaccuracy of the L1 and the 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 (.56 in Backman, 1978; .58 in Siyanova & Schmitt, 2008) and L2 speakers (.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-frequency (.74) and high-frequency (.71) collocations. Siyanova-Chanturia and Spina (2015) did not find accurate intuitions of phrasal frequency for medium-frequency and low-frequency collocations among L1 and L2 speakers, yet they concluded that both groups of participants' intuitions of high-frequency collocations correlated strongly with corpus frequency. We did not find such patterns; even for high-frequency collocations, our participants' intuitions were

still far from accurate (49.9%, 63.3%, and 56.2% for the L1 speakers, the advanced L2 speakers, and the intermediate L2 speakers, respectively). Unlike what we did in our study, Siyanova and colleagues (Siyanova & Schmitt, 2008; Siyanova-Chanturia & Spina, 2015) used nonexistent collocations as low-frequency materials created either by L2 learners or researchers. Therefore, it is unclear whether the incorporation of such stimuli might have altered their 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 assignments for 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 we reviewed previously, most studies on word frequency intuition (Backman, 1976; Balota et al., 2001; Carroll, 1971; Ringeling, 1984; Shapiro, 1969; Tryk, 1968) have 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 their choices of stimuli (e.g., the range of frequency and lexical characteristics), corpora, and tasks (e.g., the magnitude estimation task, the multiple rank order task, or simply classifying words or phrases into frequency bands). Furthermore, evidence for or against the robustness of language users' intuition of word frequency has been exclusively built upon correlation coefficients. The choice of correlation coefficients as a measure of agreement or accuracy is problematic given that (a) cutoff points for labeling correlation coefficients as weak, moderate, or strong relationship are often arbitrary and inconsistent and (b) 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 et al., 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.

This same issue with correlations arises for studies focusing on language users' intuition of phrasal frequency of multiword sequences. Therefore, to determine whether reliable intuition of word or phrasal frequency can be found, researchers should conduct more studies in which they adopt measurements of accuracy that circumvent the shortcomings of correlation coefficients. 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 estimations of frequency and of probability are 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 judgments of frequency and of probability are unavoidably error prone because they involve not only the retrieval of statistical representations but also task-irrelevant variables such as the use of strategies (judgmental heuristics). From this point of view, our finding of the inaccuracy of our L1 and L2 speakers' intuition of phrasal frequency and reports against the robustness of language users' intuition of word frequency are not surprising. Despite no previous researchers having examined language users' intuition of association strength of multiword sequences, given that L1 and L2 speakers are attuned 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 our L1 and L2 speakers' intuition of frequency and of association strength followed similar increasing patterns as corpus-based phrasal frequency and association strength increased (except for our 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 show more accurate intuitions of frequency for highly frequent collocations than for medium- and low-frequency word combinations. Our results add to the literature in that such an 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) that have claimed that linguistic knowledge—including statistical intuitions—is acquired from experience. In the case of collocations, more frequent or more associated multiword sequences are accompanied with stronger statistical representations. Last, 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 did 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 these. For intuitions of phrasal frequency, our L1 speakers had an advantage over our L2 speakers, but only for low-frequency and medium-frequency collocations. Similarly, our advanced L2 speakers' intuitions of phrasal frequency were more accurate than those of our intermediate L2 speakers, but only for high-frequency collocations. For intuitions of association strength, interestingly, our L1 speakers did not have any advantage over our L2 speakers, although we did find that our advanced L2 speakers showed more accurate intuitions than did our 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, we found in this study that, when judging the phrasal frequency and association strength of English adjective–noun collocations, the L1 and the L2 speakers not only used statistical regularities at the phrasal level but also used orthographic, phonological, and semantic characteristics of the words that constituted the word combinations. Interestingly, combining evidence from separate models using MI and log Dice as the measure of association strength, we concluded that the L1 and the L2 speakers' intuitive judgments of phrasal frequency were not affected by the corpus-based frequency bands of the collocations. Instead, we found that the L1 and the L2 speakers evaluated the degree of association between the constituent words when judging the phrasal frequency of collocations, with collocations in the high association strength band 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, the L1 and the 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, the participants perceived the words of collocations in the low-association-strength band as being associated less strongly than were

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-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. Siyanova-Chanturia and Spina (2015), however, 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; Wolter & Gyllstad, 2013; Yi, 2018; Yi et al., 2017) has shown that L1 and L2 speakers are sensitive to frequencies of multiword sequences during online tasks, processing more frequent word combinations 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 we have reported here might relate to the explicit nature of the forced-choice judgment tasks used in our study.

We also found that lexical characteristics of words that constitute collocations contribute to L1 and L2 speakers' statistical intuitions. For the estimation of phrasal frequency, we found that both L1 and L2 speakers (advanced and intermediate) make use of the orthographic (i.e., word1 length, word1/word2 orthographic neighborhood size), the phonological (i.e., word2 phonological neighborhood size), and the 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 our study, the L1 and the 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 the findings of Siyanova-Chanturia and Spina's (2015) study. Nevertheless,

Siyanova-Chanturia and Spina used noun–adjective Italian collocations instead of adjective–noun English collocations as we used in our study. 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 of which that word is a member. 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 our study, adjective–noun collocations consisting of adjectives with more orthographic neighbors and nouns with more phonological neighbors received higher ratings of phrasal frequency. Whether the choice of the measure of association strength was MI or log Dice, our L1 and our advanced L2 speakers perceived collocations that contained nouns with more orthographic neighbors and nouns that were more concrete as being used less frequently. Interestingly, such a pattern was reversed for our intermediate L2 speakers. Orthographic and phonological neighborhood size have been found to facilitate visual word recognition (Andrews, 1997; Yates et al., 2004). However, it is not clear how these variables impact language users' intuition of statistical regularities. Compared with the intuition of phrasal frequency, our L1 and our L2 speakers' judgments of association strength were less influenced by the lexical characteristics of the constituent words. The effects of word1 frequency, word1 length, and word1 orthographic neighborhood size indicated that both our L1 and our L2 speakers (advanced and intermediate) tended to rate target collocations containing adjectives that were less frequent, shorter, and with fewer orthographic neighbors as having stronger word associations. By contrast, the effect of word2 length suggested that the participants perceived target collocations with shorter nouns as words that are 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 the 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 our 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 we have reported here also echoes the debate on the holistic versus 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 some reaction-time studies have revealed.

The Relationship Between Intuitions of Phrasal Frequency and Association Strength

Our study is the first that simultaneously examined language users' intuitions of phrasal frequency and association strength of multiword sequences. As we reported previously, the L1 and the L2 speakers of our study made use of association strength when we asked them to judge collocation frequency, and they employed both phrasal frequency and association strength when judging 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 that our L1 and our 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, nonnegligible 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 access to 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 the linguistic properties of the constituent words.

Limitations and Future Directions

This study was not without limitations. As an anonymous reviewer pointed out, almost all the 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 nonhomogeneity might have altered the speakers' intuitions because qualifying adjectives and indefinite adjectives differ in the way that they describe nouns. Additionally, given that we did not consider individual learner differences other than language proficiency in our study, future

research could be conducted 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 researchers investigate 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 researchers should critically evaluate the extent to which corpora represent the input that L1 and L2 speakers receive based on the above-mentioned dimensions. Last, given that studies on intuitions of statistical regularities—especially association strength—of multiword sequences are still scarce, more studies will be needed to validate our research findings.

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 results showed that our L1 and our L2 speakers' statistical intuitions of collocations were not accurate. Furthermore, their intuitive knowledge of phrasal frequency and of association strength seemed related, despite following different cognitive processes. For the linguistic influences underlying language users' statistical intuitions, we found that knowledge of both multiword sequences and their constituent words was 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 researchers select 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 might be implicit and highly automatic.

Final revised version accepted 3 June 2022

Open Research Badges

This article has earned an Open Materials badge for making publicly available the components of the research methods needed to reproduce the reported procedure. All materials that the authors have used and have the right to share are available at https://osf.io/r9avk/ and http://www.iris-database.org. All proprietary materials have been precisely identified in the manuscript.

Notes

1 We referred to information provided by Educational Testing Services for interpreting scores from each section in TOEFL iBT on an overall proficiency scale of "below low-intermediate" to "advanced." To qualify for high-intermediate, L2 learners must attain at least a total score of 72 (18 for reading, 17 for reading, 20 for speaking, and 17 for writing); similarly, L2 learners are considered to belong to advanced if they attain at least a total score of 95 (24 for reading, 22 for listening, 25 for speaking, and 24 for writing). We retrived the information from https://www.ets.org/toefl/test-takers/ibt/scores/understanding

2 We computed MI (Yi et al., 2017) and log Dice (Öksüz et al., 2021) scores using the following mathematical formulas:

$$\mathrm{MI} = \log_2 \frac{f(xy) \times N}{f(x) \times f(y)} \tag{1}$$

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

N is the size of the corpus, and $f(xy)$, $f(x)$, 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.

3 Potential scale reduction factor, or R-hat, is a statistical index used to diagnose convergence of chains using the Markov chain Monte Carlo method. If the potential scale reduction factor is close to 1, one can conclude that chains found using the

Markov chain Monte Carlo method are well converged, and the parameter estimates are valid based on the converged chains (Gelman et al., 2013).

References

Alderson, J. C. (2007). Judging the frequency of English words. *Applied Linguistics*, *28*(3), 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*(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*(1), 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*(2), 175–186. https://doi.org/10.1080/0031383760200112

Backman, J. (1978). Subjective structures in linguistic recurrence. *Educational Reports Umeå, No. 19*. http://www.diva-portal.se/smash/get/diva2:1156056/FULLTEXT01.pdf

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, *29*(4), 639–647. https://doi.org/10.3758/BF03200465

Brysbaert, M., & New, B. (2009). Moving beyond Kučera 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*(4), 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*(3), 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*(6), 722–729. https://doi.org/10.1016/S0022-5371(71)80081-6

Davies, M. (2008). *The Corpus of Contemporary American English (COCA)*. https://www.english-corpora.org/coca

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*(5), 843–863. https://doi.org/10.1080/17470218.2012.720994

Ellis, N. C., & Ogden, D. C. (2017). Thinking about multiword constructions: Usage-based approaches to acquisition and processing. *Topics in Cognitive Science*, *9*(3), 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*. Wiley.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text & Talk*, *20*(1), 29–62. https://doi.org/10.1515/text.1.2000.20.1.29

Fletcher, W. H. (2011). *Phrases in English (PIE)*. 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*. CRC Press.

Gries, S. T., & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning*, *65*(S1), 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*(2), 1–22. https://doi.org/10.18637/jss.v033.i02

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

Kučera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. 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*(8), Article e43230. https://doi.org/10.1371/journal.pone.0043230

McCrostie, J. (2007). Investigating the accuracy of teachers' word frequency intuitions. *Regional Language Centre Journal*, *38*(1), 53–66. https://doi.org/10.1177/0033688206076158

McDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, *43*(16), 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*(1), 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* (Version 3.6.2) [Computer software]. R Foundation for Statistical Computing. http://www.R-project.org

Ringeling, T. (1984). Subjective estimations as a useful alternative to word frequency counts. *Interlanguage Studies Bulletin*, *8*(1), 59–69. https://www.jstor.org/stable/43135301

Schmitt, N., & Dunham, B. (1999). Exploring native and non-native intuitions of word frequency. *Second Language Research*, *15*(4), 389–411. https://doi.org/10.1191/026765899669633186

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia and Analgesia*, *126*(5), 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). Erlbaum.

Shapiro, B. J. (1969). Subjective estimation of relative word frequency. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 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), Article 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*(3), 429–458. https://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*(3), 533–562. https://doi.org/10.1111/lang.12125

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

Tryk, H. E. (1968). Subjective scaling of word frequency. *American Journal of Psychology*, *81*(2), 170–177. https://doi.org/10.2307/1421261

Tversky, A. K. D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/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*(3), 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*(2), 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*(3), 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., Man, K., & Maie, R. (2022a). *Data. Datasets from "Investigating L1 and L2 speaker intuitions of phrasal frequency and association strength of multiword sequences"* [Dataset]. IRIS Database, University of York, UK. https://doi.org/10.48316/9qd7-9z95

Yi, W., Man, K., & Maie, R. (2022b). *Questionnaire of statistical intuitions. Materials from "Investigating L1 and L2 speaker intuitions of phrasal frequency and association strength of multiword sequences"* [Questionnaire]. IRIS Database, University of York, UK. https://doi.org/10.48316/q0dx-yr78

Yi, W., Man, K., & Maie, R. (2022c). *Word list. Materials from "Investigating L1 and L2 speaker intuitions of phrasal frequency and association strength of multiword sequences"* [Language test]. IRIS Database, University of York, UK. https://doi.org/10.48316/tb09-j459

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 University Press.

Supporting Information

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

Accessible Summary

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 Mutual Information
Appendix S10. Accuracy of Participants' Intuition of Association Strength Measured by log Dice