The processing advantage of multiword sequences: A meta-analysis

Wei Yi [ID] and Yanlu Zhong [ID]

School of Chinese as a Second Language, Peking University, Beijing, China
Corresponding author: Wei Yi; Email: weiyisla@pku.edu.cn

Abstract

This meta-analysis synthesized 35 English studies (130 effect sizes, $N = 1,981$) that employed online tasks to investigate the processing of multiword sequences (MWSs). We examined (a) to what extent MWSs enjoy a processing advantage over novel word combinations; (b) how such a processing advantage is moderated by statistical regularities (i.e., phrasal frequency, association strength), MWS type, and explicitness of experimental tasks; and (c) whether such moderating patterns differ between L1 speakers and L2 speakers. The results confirmed the processing advantage for most subtypes of MWSs, with effect sizes ranging from small to medium. For L1 speakers and L2 speakers, the processing advantage of MWSs was found across the continuum of phrasal frequency and association strength and varied. Interestingly, task explicitness moderated the processing advantage of MWSs but only for L2 speakers. Taken together, our results shed light on the understanding of MWSs as well as directions for future research.

Introduction

Like single words, multiword sequences (MWSs) are said to be building blocks for language acquisition and processing (Christiansen & Arnon, 2017). Over the past decades, a growing number of studies have reported the processing advantage of MWSs over novel word combinations (i.e., word combinations generated creatively) for L1 speakers and L2 speakers (Conklin & Carrol, 2021; Jiang & Nekrasova, 2007; Siyanova-Chanturia et al., 2011). However, individual studies vary substantially in terms of experimental design, which leads to inconsistencies in research findings and leaves unclear (a) to what extent MWSs enjoy a processing advantage over novel word combinations, (b) which factors moderate the processing advantage of MWSs, and (c) whether L1 speakers and L2 speakers differ when processing MWSs. To address these issues, we systematically synthesized studies that employed online tasks to explore the processing advantage of MWSs. We incorporated statistical regularities (i.e., phrasal frequency and association strength), MWS type, and task explicitness as the moderators and compared their effects on the processing advantage of MWSs for L1 speakers and L2 speakers.

Literature review

MWSs and their processing advantage

Language can be represented at multiple levels. From the perspective of usage-based approaches (Christiansen & Arnon, 2017; Goldberg, 1995), MWSs—which are word strings that co-occur more often than by chance—are integral building blocks of language. MWSs cover a wide variety of linguistic phenomena, including but not limited to idioms (e.g., *kick the bucket*), speech formulae (e.g., *what's up*), phrasal verbs (e.g., *take off*), binomials (e.g., *knife and fork*), collocations (e.g., *make progress*), and lexical bundles (e.g., *is one of the*). Corpus studies have found that MWSs are highly frequent and widely used in language (Biber, 2009). Moreover, they facilitate the development of fluency and nativelikeness for language speakers (Wray & Perkins, 2000). Many researchers (Wray, 2002) believe that MWSs—especially formulaic (i.e., highly conventionalized) ones—are prefabricated chunks stored in the memory. Following this, they may enjoy a processing advantage over novel word combinations and free up cognitive resources by reducing the time pressure during language processing (Christiansen & Chater, 2016).

Over the past decades, the processing advantage of MWSs has attracted an increasing amount of attention. A plethora of studies have found that various types of MWSs are processed significantly faster than are novel word strings, with similar results being reported in both children (Bannard & Matthews, 2008) and adults (Arnon & Snider, 2010), for both L1 speakers (Siyanova-Chanturia et al., 2011) and L2 speakers (Wolter & Yamashita, 2018), using both online (Sonbul, 2015; Tremblay & Baayen, 2010) and offline tasks (Sonbul, 2015). However, given that the experimental design varies drastically across studies, it remains difficult to estimate to what extent MWSs are processed significantly faster than are novel word combinations.

Variables that may moderate the processing advantage of MWSs

Considerable evidence supports the processing advantage of MWSs over novel word combinations. However, a comprehensive understanding of the variables that could potentially moderate this advantage and the mechanisms through which they operate is still lacking. In this section, we will review three variables that seem most worth considering.

Statistical regularities

Language input is not random. Instead, it is characterized by underlying statistical regularities. Language users can capture distributional information and make use of it for phonological learning, word segmentation, and syntactic learning (for reviews, see Ellis, 2002; Saffran, 2003). Usage-based approaches (Christiansen & Arnon, 2017; Goldberg, 1995) hold that language acquisition is exemplar based and shaped by repeated language use. By tracking statistical regularities, language users can establish and modify mental representations of linguistic units at various levels.

Recent research suggests that two types of statistical regularities play crucial roles in the processing of MWSs—namely, phrasal frequency and association strength (Gries & Ellis, 2015; Yi, 2018; Yi et al., 2017). Phrasal frequency indicates how likely a word string is to be experienced by language users (Gries & Ellis, 2015). In contrast, association strength evaluates the co-occurring probability of words that constitute MWSs and how likely language users can predict the words following or preceding another word in a

sequence (Gablasova et al., 2017). Phrasal frequency and association strength are usually computed based on corpora data, with the latter measured by metrics such as mutual information (MI), transitional probability, T-score, delta P, and log Dice (for reviews, see Gries, 2022; Gries & Ellis, 2015; Yi, et al., 2023). Recent studies have demonstrated that L1 speakers and L2 speakers are sensitive to both phrasal frequency (Arnon & Snider, 2010; Tremblay & Baayen, 2010; Yi, 2018; Yi et al., 2017) and association strength (Gyllstad & Wolter, 2016; Öksüz et al., 2021; Yi, 2018; Yi et al., 2017) when processing MWSs. Specifically, higher frequency MWSs are processed significantly faster than are lower frequency ones (e.g., Durrant, 2008; Kim & Kim, 2012; Sonbul, 2015). Similarly, MWSs consisting of words that are more closely associated also tend to be processed more efficiently than those constituted by loosely associated words (Öksüz et al., 2021; Yi, et al., 2017).

The effects of statistical regularities on the processing of MWSs have been observed across the entire range (Arnon & Snider, 2010; Yi, 2018; Yi et al., 2017). Nevertheless, methodologically, researchers often define and select MWSs based on certain thresholds (e.g., occurring at least 10 times per million words; Biber et al., 1999). Consequently, examining statistical regularities as moderators of the processing advantage of MWSs would enable us to determine whether this processing advantage is restricted to specific statistical profiles. Moreover, it would provide insight into whether variations in the processing advantage of MWSs can be attributed to variations in statistical regularities, thereby enhancing our understanding of whether language users use statistical information to process MWSs more efficiently.

MWS Type

MWS is an umbrella term that encompasses a wide range of larger-than-word units. MWSs are not as homogenous as one might think; instead, subtypes of MWSs vary drastically in terms of structural, semantic, and syntactic characteristics. For instance, whereas MWSs such as idioms, collocations, binomials, and phrasal verbs are structurally complete and self-contained, lexical bundles are structurally incomplete and often span syntactic boundaries (Jeong & Jiang, 2019). Idioms are semantically figurative and noncompositional, with constituent words contributing little to the meaning of the whole word string (Carrol & Conklin, 2020). In contrast, lexical bundles are semantically transparent and compositional, whereas collocations (e.g., *build a career* vs. *build a house*) and phrasal verbs (e.g., *rise up* vs. *heat up*) can be interpreted both literally and figuratively.

A wealth of studies has demonstrated the processing advantage of MWSs, including idioms (Siyanova-Chanturia et al., 2011), lexical bundles (Tremblay et al., 2011), binomials (Carrol & Conklin, 2020), collocations (Wolter & Yamashita, 2018), and phrasal verbs (Cappelle et al., 2010; Hanna et al., 2017). However, little research has examined whether such a processing advantage varied significantly across subcategories of MWSs. Columbus and Wood (2010) compared the processing of idioms, lexical bundles, and collocations using an eye-tracking reading task. The results showed that L1 speakers of English read all three types of MWSs more quickly than control items, and idioms were processed faster than were lexical bundles and collocations. Jeong and Jiang (2019) examined the processing of structurally complete formulaic expressions (e.g., *for example*) and structurally incomplete lexical bundles (e.g., *is one of the most*) using a word-monitoring task. For both L1 speakers and L2 speakers of English, a processing advantage was found for formulaic sequences but not for lexical bundles. Gyllstad and Wolter (2016) compared the processing of novel word

combinations (e.g., *sing a song*) and collocations (e.g., *keep a secret*) by L1 speakers and L2 speakers, using a semantic judgment task. The results showed a processing disadvantage for semitransparent collocations relative to semantically fully transparent word combinations, indicating that semantic transparency plays an important role in the processing of MWSs. By incorporating MWS type as a moderator, the extent of the processing advantage for subcategories of MWSs can be unveiled. This would enhance our comprehension of the multifaceted nature of multiword expressions, identify the subcategories of MWSs that pose greater challenges for L2 speakers to acquire, and contribute to the development and refinement of language processing models.

Explicitness of experimental tasks

Both online and offline experimental tasks—have been used to study the processing of MWSs. Different from offline tasks, online tasks are performed under significant time pressure and are more likely to tap into cognitive processes involved during the processing of MWSs than are offline tasks (Siyanova-Chanturia, 2015). Following this, it is not uncommon to see that result patterns with respect to language processing revealed by offline tasks may not be consistent with those obtained from online tasks (Pellicer-Sánchez et al., 2022; Sonbul, 2015). Recent studies (Suzuki & DeKeyser, 2017) in the field of second language acquisition have shown that experimental tasks also vary in the degree of task explicitness. Task explicitness refers to the extent to which a task requires explicit or implicit knowledge. For instance, Suzuki and DeKeyser (2015, 2017) concluded that eye tracking (i.e., visual-world paradigm), self-paced reading, and word-monitoring tasks measure implicit knowledge, whereas timed grammaticality judgment and elicited imitation tap into automatized explicit knowledge. Yi (2018) adopted a phrasal acceptability judgment task (PJT) to examine the processing of English collocations by L1 speakers and L2 speakers. Based on the correlational relationships between PJT performance and language aptitudes, he suggested that L1 speakers might process collocations implicitly, whereas L2 speakers might process collocations explicitly. To the best of our knowledge, little research has investigated whether task explicitness affects the processing advantage of MWSs. By incorporating task explicitness as a moderator in a meta-analysis, we could explore whether the processing advantage of MWSs is task dependent, thus shedding light on the cognitive mechanism underlying MWSs processing as well as the selection of experimental tasks for future studies.

Potential differences between L1 speakers and L2 speakers

Numerous studies have found that both L1 speakers (Carrol & Conklin, 2020) and L2 speakers (Jeong & Jiang, 2019; Wolter & Yamashita, 2018) process MWSs significantly faster than they do novel word combinations. However, research findings have not always been consistent and the magnitude of the processing advantage of MWSs varies among studies. For example, although the processing advantage of idioms is well established for L1 speakers (Siyanova-Chanturia et al., 2011), some studies (Carrol & Conklin, 2014) did not replicate such patterns for L2 speakers. In addition, among those studies that compared L1 speakers and L2 speakers, some found the processing advantage of MWSs—such as idioms (Siyanova-Chanturia et al., 2011) and lexical bundles (Jeong & Jiang, 2019)—for L1 speakers but not for L2 speakers. Such inconsistencies create difficulties in determining whether the processing advantage of MWS varies in degree between L1 speakers and L2 speakers and whether subtypes of MWSs

may or may not enjoy a processing advantage for L1 speakers and L2 speakers. Regarding language users' sensitivity to statistical regularities when processing MWSs, Ellis et al. (2008) suggest that L1 speakers' processing of MWSs could be primarily influenced by association strength (measured by MI), whereas L2 speakers might exhibit greater sensitivity to phrasal frequency. Considering the significant disparities in learning context and language experience between L1 speakers and L2 speakers, it is crucial to explore whether the moderating effects of statistical regularities and task explicitness on the processing advantage of MWSs differ between these two groups.

The current study

As our review has shown, although many studies have reported the processing advantage of MWSs, due to methodological variations across studies, it remains unknown (a) to what extent MWSs are processed faster than novel word combinations; (b) how the processing advantage of MWSs is moderated by statistical regularities, type of MWSs, and task explicitness; and (c) whether such moderating patterns differ between L1 speakers and L2 speakers. Meta-analysis allows the examination of variables that were not the focus of individual studies. To address these research questions (RQs), we conducted a meta-analysis and synthesized studies that examined the processing advantage of MWSs using online experimental tasks.

Method
Literature search

We used the following databases to identify studies to include in this meta-analysis: Education Resources Information Center (ERIC), Google, Google Scholar, Linguistics and Language Behavior Abstracts (LLBA), Oxford Bibliographies, ProQuest Dissertations and Theses, PsycINFO, VARGA (Vocabulary Acquisition Research Group Archive), and Web of Science. We chose to start our meta-analysis from the year 2000 because there was a growing interest in the processing advantage of MWSs during that period and Wray's work in 2002 (Wray, 2002) made a significant contribution to the field and provided the basis for future studies on the topic. We searched these databases for abstracts published from January 2000 to January 2022 using the following keywords: binomials, collocations, formulaic language, formulaic sequences, formulaic speech, frozen phrases, idioms, lexical bundles, multiword expressions, multiword sequences, multiword units, prefabricated language, prefabricated patterns, phrasal verbs, word combinations, and word sequences. Additionally, we conducted a forward citation search on reports containing the above terms (see Appendix S1 for the PRISMA flow diagram for the inclusion of studies).

Inclusion and exclusion criteria

The following criteria were employed to determine which studies to include in this meta-analysis.

1) Significant discrepancies exist in the characteristics of MWS across languages. To ensure the comparability of research findings, we included only studies that were written in English and used English stimuli.

2) We excluded literature reviews and empirical studies that focused on the instruction and acquisition of novel multiword expressions.
3) We included studies that investigated the processing of MWSs using online tasks and excluded those that used only offline tasks.
4) We included studies that adopted reading time and reaction time (RT) as the outcome variable, following previous studies (Avery & Marsden, 2019).
5) We excluded those that only reported accuracies of experimental trials (accuracy data are less replicable and difficult to interpret, Jiang, 2012) or neural responses (e.g., EEG, ERP, fMRI).
6) We excluded studies that used production tasks (e.g., word naming), given that production tasks tap into cognitive processes that are different from those underlying perception tasks (e.g., lexical decision), Specifically, language users encode input into multiword units and pass them to a higher level of linguistic representation during perception tasks (Christiansen, 2016; Jiang, 2012), whereas they retrieve ready-made units appropriate for conversation and piece them together in the opposite direction during production tasks (Arnon & Priva, 2013).
7) We excluded studies for which full texts were not available.
8) We excluded studies that did not provide enough information for calculating effect sizes (i.e., mean, *SD*, number of participants or items).
9) Thirty-five studies (130 effect sizes, 1,981 participants) met all the criteria. These studies comprised 29 journal articles, two doctoral dissertations, three book chapters, and one conference proceeding (see Appendix S2).

Coding

We coded the included studies for study identifiers, moderator variables, language background, and descriptive statistics for calculating effect sizes (see Appendix S3 for the coding sheet).

Statistical regularities

We coded phrasal frequency and association strength of MWSs separately. Specifically, we chose MI, forward delta *P*, and backward delta *P* as measures of association strength for the following reasons. First, many studies have documented that L1 speakers and L2 speakers are sensitive to the association strength of MWSs measured by MI and delta *P* (Gries, 2013; Öksüz et al., 2021; Yi, 2018). Second, MI assumes the association between words as nondirectional, whereas delta *P* can evaluate the degree to which the first or the last word in a sequence is predicted by other words in a forward or backward direction (Gries, 2013). Therefore, using both MI and delta *P*s should allow us to test the directionality of association strength for MWSs (for the calculation of MI and delta *P*s, see Appendix S4).

Studies included in this meta-analysis varied in their choice of corpora when retrieving phrasal frequency and association strength for MWSs. To ensure the validity of statistical regularities, we followed the practice of Lindstromberg and Eyckmans (2020) and requeried the corpus statistics for MWSs in each study from the British

National Corpus (BNC). The selection of BNC over other corpora, such as COCA (Contemporary Corpus of American English), was predicated on several factors. First, a greater number of studies (N = 18) have reported statistical regularities derived from BNC than reported from COCA (N = 9). Second, despite potential disparities between British and American English, prior research (Sonbul, 2015; Yi et al., 2023) has shown that corpus data from BNC and COCA are similar and highly correlated. Following previous studies (Siyanova-Chanturia & Spina, 2015; Yi et al., 2023), we first calculated the mean phrasal frequency, mean MI, and mean delta Ps based on statistics extracted from the BNC. We then ranked the means from lowest to highest and used the lower, median, and upper quartiles as cutoff points to split the statistical regularities into low, medium, high, and very high bins as shown in Table 1 (see Appendix S5 for more details).

MWS type

MWSs originally labeled in the included studies consisted of collocations (15 studies), idioms (11 studies), lexical bundles (4 studies), binominals (4 studies), formulaic sequences (2 studies), and multiword sequences/units (2 studies). After double-checking the word strings in the original research, we relabeled the stimuli in some studies and coded MWSs in all included studies into five categories: collocations (15 studies, 64 effect sizes), idioms (11 studies, 31 effect sizes), lexical bundles (6 studies, 17 effect sizes), binominals (4 studies, 12 effect sizes), and phrasal verbs (1 study, 6 effect sizes). Details about the grouping and definitions of each type of MWSs are provided in Appendix S6.

Task explicitness

The experimental tasks that were employed in the included studies comprised eye-tracking reading (ET; 13 studies, 43 effect sizes), grammaticality judgment (GJT; 1 study, 4 effect sizes), lexical decision (LDT; 8 studies, 32 effect sizes), phrasal acceptability judgment (PJT; 9 studies, 37 effect sizes), self-paced reading (SPR; 3 studies, 10 effect sizes), and word monitoring (WM; 1 study, 4 effect sizes; see Appendix S7 for the descriptions of tasks). Previous studies (Ellis, 2005; Suzuki & DeKeyser, 2015, 2017) operationalized implicit and explicit knowledge tapped into by tasks based on the degree of awareness, time available, focus of attention, and meta-linguistic knowledge, among which awareness is regarded as the core criterion. Given that awareness is subjective and extremely difficult to measure in real time, we dropped the awareness criterion and coded the explicitness of tasks based on time available, focus of attention, and metalinguistic knowledge.

Language users tend to be more aware of their knowledge employed in tasks carried out under less time pressure (Suzuki & DeKeyser, 2017). Following this, tasks were coded as most explicit when participants completed them without time limits (i.e., at their own pace), whereas tasks were coded as less or least explicit when participants completed them as quickly as possible or with time limits.[1] We operationalized the focus of attention for each task based on the availability of contextual information when

---

[1]Most eye-tracking studies included in our meta-analysis required participants to read passages at their own pace (i.e., without any time pressure). According to our coding scheme, these studies were coded as most explicit in terms of time pressure. Two eye-tracking studies (Kyriacou et al., 2021; Siyanova-Chanturia et al.,

Table 1. Average statistical regularities of MWSs across bins

Bins	Phrasal frequency			Association strength								
				MI			Forward delta P			Backward delta P		
	n	MWSs	Control	n	MWSs	Control	n	MWSs	Control	n	MWSs	Control
Low	24	0.102	0.025	24	3.340	1.660	22	0.005	0.003	22	0.002	0.000
Medium	29	0.366	0.440	35	7.080	7.570	29	0.026	0.023	34	0.014	0.015
High	36	1.410	1.200	27	9.560	9.270	32	0.088	0.072	26	0.078	0.139
Very high	30	11.400	4.170	33	16.000	13.800	34	0.397	0.169	35	0.475	NA

Note. n represents the number of effect sizes in each bin. *NA* indicates that there was no data in the bin. Control refers to novel word combinations. Phrasal frequencies were transformed to number of occurrences per million words. Missing values were found for phrasal frequency (11), MI (11), forward delta P (13), and backward delta P (13).

processing MWSs given that language users are more likely to focus on meaning rather than on form when more contextual support is provided (Long, 1991; Suzuki & DeKeyser, 2017). Specifically, tasks (e.g., PJT, LDT) were coded as most explicit when MWSs were presented to participants without any contextual support, whereas tasks (e.g., SPR) were coded as less explicit when MWSs were embedded in individual sentences. Tasks (e.g., ET) were coded as least explicit when MWSs were presented to participants in paragraphs. Regarding metalinguistic knowledge, tasks (e.g., GJT) were coded as most explicit when they involved the use of analytic linguistic rules, whereas tasks were coded as less (e.g., LDT, PJT) or least explicit (e.g., SPR, ET) when they required minimum or no use of metalinguistic knowledge. Based on the scheme presented above, we assigned 1, 2, or 3 points to each category of the three dimensions to indicate increasing levels of task explicitness. Then, we rated each task on these dimensions and added up the points (range: 4–8; see Table 2;[2] see Appendix S3 for the coding of task explicitness in each study). We separated the points into two groups, where tasks receiving 4 and 5 points were considered as low explicitness (45 effect sizes), and tasks receiving 6, 7 and 8 points were considered as high explicitness (85 effect sizes).

Language background

To examine whether L2 speakers differ from L1 speakers during the online processing of MWSs, we coded participants' language backgrounds as L1 (76 effect sizes) or L2 (54 effect sizes). We considered coding L2 proficiency into beginning, intermediate, and advanced levels, yet this idea was abandoned because the majority of included

Table 2. The coding of task explicitness

Task	n	Time available	Focus of attention	Metalinguistic knowledge	Total points	Task explicitness
PJT	9	1–3	3	2	6–8	low/high
WM	1	1–3	3	2	7	high
LDT	8	1–3	3	2	6–8	low/high
SPR	3	1–3	2	1	5–6	low
ET	13	2–3	1	1	4–5	low
GJT	1	2	3	3	8	high

Note. n = the number of studies.

2011) asked participants to read sentences as quickly as possible, and they were coded as less explicit in terms of time pressure.

[2]Based on our coding scheme, for an experimental task that received a total score of 9, it should (a) require participants to perform the task without any time pressure (at their own pace), (b) explicitly draw participants' attention to the linguistic form (i.e., MWS), and (c) fully involve the use of metalinguistic knowledge (e.g., grammatical rules). However, among the included studies in this meta-analysis, not a single task (PJT, WM, LDT, SPR, ET, GJT) had all three of these features. In fact, among the included studies in our meta-analysis, there was also not a single study that received a task explicitness score of 3 (i.e., with time pressure, focus of attention to the meaning of sentences/texts, without the involvement of metalinguistic knowledge).

studies (97.1%) recruited advanced-level L2 participants and their measurement of L2 proficiency varied drastically across studies.

Coding procedure

The authors of this meta-analysis coded all the included studies separately using the developed coding scheme. We calculated the intercoder reliability using Cohen's kappa test and found the agreement rate was κ = .997 (see Appendix S8 for Cohen's kappa for each coded moderator).

Data analysis

Calculation of effect size

To calculate effect sizes, Cohen's *d*, the standardized mean differences of RTs from a study, were estimated and then converted to Hedges's *g* by multiplying the correction factor *J* (Borenstein et al., 2009) to address biases arising from small sample sizes. The calculation of all the measures is described in Appendix S9. For eye-tracking studies, we used total reading time to calculate effect sizes, given this measure sums all fixations made within a region of interest (Liversedge et al., 1998) and reflects the integration of information during language processing.

Analysis procedure

We conducted the analyses in the R statistical environment (version 4.1.2, R Core Team, 2021) using the *metafor* package (version 3.0.2, Viechtbauer, 2010) and the *clubSandwich* package (version 0.5.5, Pustejovsky, 2022). To deal with potential Type I errors due to dependencies among effect sizes, we built three-level meta-regression models (e.g., Yanagisawa & Webb, 2021), which encompassed the sampling variance of each effect size (Level 1), the variance between effect sizes within the same study (Level 2), and the variance between studies (Level 3). Three-level meta-regression models can be viewed as an expansion of the conventional random-effects model (Yanagisawa & Webb, 2021). In addition to three-level regressions, we applied cluster-robust variance estimation (Hedges et al., 2010) with small sample adjustments (Tipton & Pustejovsky, 2015) to control biases due to the dependency of effect sizes (Pustejovsky, 2015). We set the significance level at α = .05. However, according to Greenland et al. (2016), it is advisable to perceive the *p* value as a continuous measure rather than a threshold. Therefore, following the practice of previous studies (e.g., Yanagisawa & Webb, 2021), *p* values lower than .10 were interpreted as marginal significance, indicating a trend effect.

To estimate the overall processing advantage of MWSs over novel word combinations (RQ$_1$), we built a three-level model and calculated the weighted average effect size using the L1-L2 mixture data. Given the potential heterogeneity between L1 speakers and L2 speakers, we also split the data and calculated weighted average effect sizes for L1 speakers and L2 speakers, respectively. To evaluate the moderating effects of statistical regularities, MWS type, and task explicitness (RQ$_2$) on the processing advantage of MWSs and to examine potential differences between L1 speakers and L2 speakers in terms of such moderating effects (RQ$_3$), we followed the practice of previous studies (e.g., Yanagisawa & Webb, 2021) and conducted separate meta-regressions with these moderators for L1 speakers and L2 speakers separately. Additionally, we tested the

interaction between language background (L1 vs. L2) and the moderators using the L1-L2 mixture data to examine whether group comparisons for each moderator differ significantly between L1 speakers and L2 speakers. For RQ$_2$ and RQ$_3$, given control phrases in the included studies varied in phrasal frequency and association strength, we incorporated the statistical regularities of control phrases as covariates. In addition, we performed multiple comparisons by changing the reference levels (e.g., Yanagisawa et al., 2020) to examine how the processing advantage of MWSs varied between levels of each moderator. Multiple comparisons between levels in each moderator were indicated by unstandardized coefficient estimates and *p* values. Main effects of the moderators in three-level models were examined by a conservative type of Wald test (i.e., HTZ test; Tipton & Pustejovsky, 2015) from which *F* statistics were obtained. Prior to the aggregation of effect sizes and moderator analyses, we conducted three analyses to assess the potential influence of publication bias on our data set: fail-safe *N*, Orwin's fail-safe *N*, and the trim-and-fill method (Borenstein et al., 2009). All three measures indicated that no publication bias was present (see Appendix S10 for more details). Additionally, we conducted four sensitivity analyses to examine whether our meta-analysis results could be replicated under different scenarios (for more details, see Appendix S11).

Results

After aggregating 130 effect sizes from 35 studies, we found that the weighted average effect size was significant (Hedges's *g* = 0.417; 95% CI = [0.276, 0.558], *p* < .001), suggesting that MWSs are processed significantly faster than novel word combinations. Regarding the heterogeneity of effect sizes, a significant *Q* statistic (*Q* = 451.74, *p* < .001) was found, indicating heterogeneity of effect sizes across the studies. The estimated variance components were $\tau^2 = 0.152$ between studies and $\tau^2 < 0.001$ within studies. The results of I^2 revealed that 70.91% of the total variance could be attributed to between-study heterogeneity ($I^2 = 70.91\%$), whereas almost no variance could be attributed to within-study heterogeneity ($I^2 < 0.01\%$). In addition, the prediction interval was [-0.77, 1.72] (see Appendix S12 for the forest plot).

By using language background as a moderator, the results showed that language users process MWSs significantly faster than novel word combinations (for L1 speakers: Hedges's *g* = 0.475, 95% CI = [0.313, 0.636], *p* < .001; for L2 speakers: Hedges's *g* = 0.348, 95% CI = [0.220, 0.475], *p* < .001). The Wald test showed that the processing advantage of MWSs for L1 speakers was significantly greater than that for L2 speakers (*F (1, 16.3)* = 11.90, *p* = .003).

The results of the separate moderator analyses after splitting the data for L1 speakers and L2 speakers are presented in Table 3 and Table 4, respectively (see Appendix S13 for the results of the moderator analyses based on the L1-L2 mixture data). For both L1 speakers and L2 speakers, the processing advantages of MWSs across all the bins of phrasal frequency and association strength (see Figure 1) were significantly or marginally significantly greater than 0.

The processing advantages of MWSs across frequency bins ranged from 0.407 to 0.862 for L1 speakers and from 0.148 to 0.631 for L2 speakers. Wald tests revealed a marginally significant main effect of frequency for L1 speakers (*F* = 4.30, *p* = .079) and L2 speakers (*F* = 4.51, *p* = .062). Subsequent multiple comparisons revealed that the processing advantage of MWSs in the very-high- (L1 speakers: *b* = 0.455, 95% CI = [0.170, 0.740], *p* = .029; L2 speakers: *b* = 0.483, 95% CI = [0.239, 0.727], *p* = .006) and

Table 3. Results of moderator analyses based on L1 data

Moderator	k	n	Mean effect size for L1 g [95% CI]	p	Regression for L1 b [95% CI]	p	Wald test F	p
Frequency (Total R^2 = .050, Between-study R^2 = .050, Within-study R^2 = .303)							4.30	.079
Low (Intercept)	11	16(22.2%)	0.407[0.230, 0.584]	<.001	0.407[0.230, 0.584]	<.001		
Medium	10	16(22.2%)	0.455[0.236, 0.674]	.001	0.049[-0.107, 0.204]	.570		
High	12	23(31.9%)	0.677[0.396, 0.958]	<.001	0.270[0.086, 0.455]	.040		
Very high	10	17(23.6%)	0.862[0.502, 1.222]	.003	0.455[0.170, 0.740]	.029		
MI (Total R^2 = .078, Between-study R^2 = .078, Within-study R^2 = .086)							4.98	.051
Low (Intercept)	8	16(22.2%)	0.500[0.243, 0.758]	.003	0.500[0.243, 0.758]	.003		
Medium	10	20(27.8%)	0.760[0.517, 1.003]	<.001	0.260[0.126, 0.394]	.025		
High	13	17(23.6%)	0.567[0.283, 0.852]	.001	0.067[-0.090, 0.224]	.456		
Very high	13	19(26.4%)	0.481[0.290, 0.672]	<.001	-0.019[-0.269, 0.230]	.883		
Forward delta P (Total R^2 = .159, Between-study R^2 = .159, Within-study R^2 = .393)							4.18	.073
Low (Intercept)	8	14(20.0%)	0.320[0.127, 0.514]	.007	0.320[0.127, 0.514]	.007		
Medium	10	19(27.1%)	0.684[0.433, 0.935]	<.001	0.364[0.180, 0.547]	.027		
High	13	18(25.7%)	0.686[0.381, 0.991]	<.001	0.366[0.111, 0.621]	.019		
Very high	15	19(27.1%)	0.679[0.449, 0.909]	<.001	0.359[0.118, 0.599]	.015		
Backward delta P (Total R^2 = .064, Between-study R^2 = .064, Within-study R^2 = .465)							3.17	.096
Low (Intercept)	7	14(20.0%)	0.438[0.136, 0.739]	.019	0.438[0.136, 0.739]	.019		
Medium	12	20(28.6%)	0.662[0.379, 0.945]	<.001	0.225[0.081, 0.368]	.046		
High	11	16(22.9%)	0.623[0.325, 0.921]	.001	0.185[-0.014, 0.384]	.109		
Very high	13	20(28.6%)	0.538[0.285, 0.791]	.003	0.101[-0.234, 0.436]	.566		
MWS type (Total R^2 = .282, Between-study R^2 = .282, Within-study R^2 = .500)							141	<.001
Collocations (Intercept)	14	38(50.0%)	0.629[0.327, 0.930]	.001	0.629[0.327, 0.930]	.001		
Idioms	11	19(25.6%)	0.383[0.161, 0.604]	.007	-0.246[-0.620, 0.128]	.221		
Lexical bundles	4	7(9.2%)	0.258[0.138, 0.378]	.025	-0.370[-0.695, -0.046]	.077		
Binominals	4	9(11.8%)	0.432[0.069, 0.796]	.110	-0.196[-0.598, 0.205]	.401		
Phrasal verbs	1	3(3.9%)	3.309[3.151, 3.468]	.016	2.681[2.341, 3.021]	.001		
Task explicitness (Total R^2 = .000, Between-study R^2 = .000, Within-study R^2 = .000)							1.06	.314
Low (Intercept)	13	29(38.2%)	0.419[0.184, 0.653]	.004	0.419[0.184, 0.653]	.004		
High	19	47(61.8%)	0.612[0.328, 0.895]	<.001	0.193[-0.175, 0.561]	.314		

Note. k = the number of studies; *n* = the number of effect sizes; *g* = Hedges's *g*; *b* = unstandardized coefficient; CI = confidence interval; ref = reference level. Missing values were present when coding the moderators. The percentages of studies were calculated after excluding missing values.

Table 4. Results of moderator analyses based on L2 data

Moderator	k	n	Mean effect size for L2		Regression for L2		Wald test	
			g [95% CI]	p	b [95% CI]	p	F	p
Frequency (Total R^2 = .108, Between-study R^2 = .108, Within-study R^2 = .000)							4.51	.062
Low (Intercept)	7	8(17.0%)	0.148[0.053, 0.243]	.021	0.148[0.053, 0.243]	.021		
Medium	8	13(27.7%)	0.441[0.205, 0.677]	.006	0.293[0.097, 0.489]	.029		
High	9	13(27.7%)	0.500[0.226, 0.774]	.008	0.351[0.118, 0.584]	.015		
Very high	7	13(27.7%)	0.631[0.379, 0.883]	.010	0.483[0.239, 0.727]	.006		
MI (Total R^2 = .069, Between-study R^2 = .069, Within-study R^2 = .247)							2.55	.159
Low (Intercept)	5	8(17.0%)	0.313[0.132, 0.494]	.022	0.313[0.132, 0.494]	.022		
Medium	7	15(31.9%)	0.648[0.336, 0.960]	.008	0.335[0.111, 0.559]	.033		
High	8	10(21.3%)	0.431[0.149, 0.714]	.018	0.119[-0.073, 0.310]	.282		
Very high	11	14(29.8%)	0.160[0.083, 0.237]	.004	-0.153[-0.336, 0.031]	.133		
Forward delta P (Total R^2 = .300, Between-study R^2 = .300, Within-study R^2 = .000)							1.93	.235
Low (Intercept)	5	8(17.0%)	0.273[0.161, 0.386]	.006	0.273[0.161, 0.386]	.006		
Medium	5	10(21.3%)	0.669[0.311, 1.027]	.015	0.396[0.111, 0.681]	.038		
High	10	14(29.8%)	0.513[0.238, 0.787]	.004	0.239[0.015, 0.463]	.088		
Very high	13	15(31.9%)	0.268[0.125, 0.411]	.009	-0.005[-0.176, 0.166]	.954		
Backward delta P (Total R^2 = .112, Between-study R^2 = .112, Within-study R^2 = .304)							1.56	.303
Low (Intercept)	5	8(17.0%)	0.493[0.048, 0.937]	.082	0.493[0.048, 0.937]	.082		
Medium	8	14(29.8%)	0.618[0.322, 0.914]	.004	0.125[-0.160, 0.411]	.460		
High	6	10(21.3%)	0.367[0.074, 0.661]	.051	-0.125[-0.618, 0.367]	.636		
Very high	11	15(31.9%)	0.253[0.073, 0.434]	.038	-0.239[-0.709, 0.231]	.343		
MWS type (Total R^2 = .114, Between-study R^2 = .114, Within-study R^2 = .494)							9.23	.030
Collocations (Intercept)	10	26(48.1%)	0.530[0.199, 0.861]	.013	0.530[0.199, 0.861]	.013		
Idioms	8	12(22.2%)	0.073[-0.035, 0.181]	.230	-0.457[-0.805, -0.109]	.022		
Lexical bundles	6	10(18.5%)	0.308[0.215, 0.400]	.001	-0.222[-0.566, 0.121]	.232		
Binominals	2	3(5.6%)	0.138[-0.058, 0.333]	.400	-0.392[-0.777, -0.008]	.234		
Phrasal verbs	1	3(5.6%)	0.605[0.520, 0.691]	.046	0.075[-0.266, 0.417]	.676		
Task explicitness (Total R^2 = .106, Between-study R^2 =.106, Within-study R^2 = .329)							6.84	.020
Low (Intercept)	9	16(29.6%)	0.124[0.017, 0.230]	.054	0.124[0.017, 0.230]	.054		
High	18	38(70.4%)	0.426[0.226, 0.625]	<.001	0.302[0.076, 0.528]	.020		

Note. k = the number of studies; n = the number of effect sizes; g = Hedges's g; b = unstandardized coefficient; CI = confidence interval; ref = reference level. Missing values were present when coding the moderators. The percentages of studies were calculated after excluding missing values.
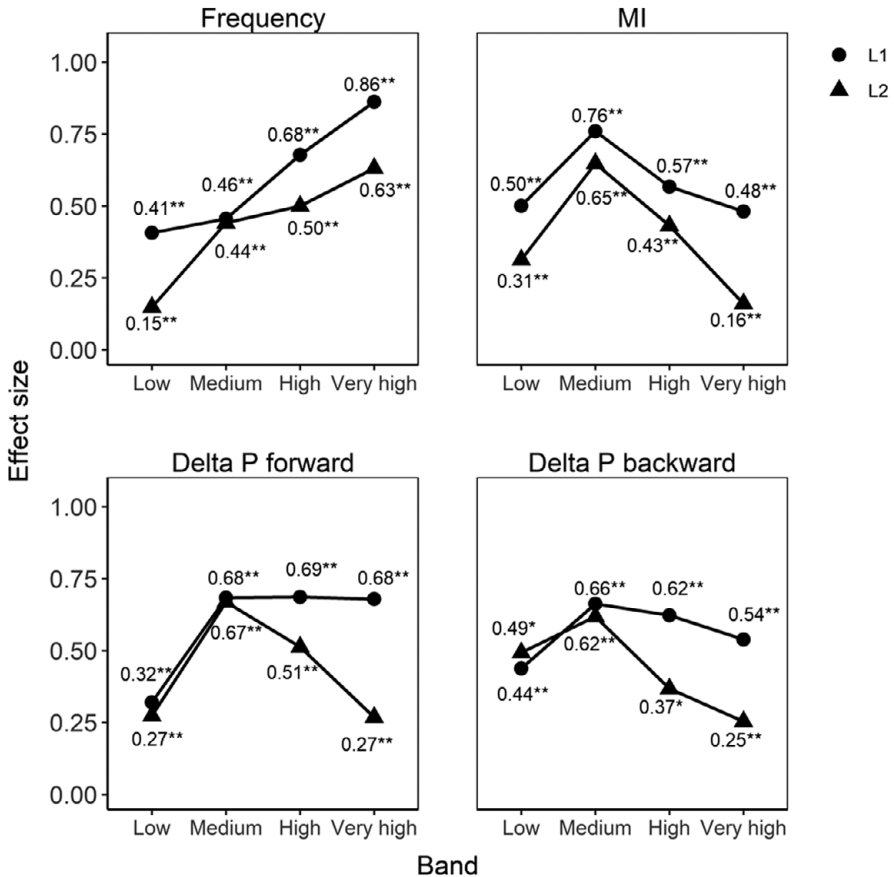
The processing advantage of multiword sequences: A meta-analysis

Figure 1. . Processing advantages of MWSs broken down by statistical regularities and language background. *$p < .10$; **$p < .05$.

high-frequency bins (L1 speakers: $b = 0.270$, 95% CI = [0.086, 0.455], $p = .040$; L2 speakers: $b = 0.351$, 95% CI = [0.118, 0.584], $p = .015$) were significantly greater than that in the low-frequency bin (see Appendix S14 for multiple comparison results). We also found differences between L1 speakers and L2 speakers. Specifically, L1 speakers in the medium-frequency bin did not show a significant advantage in processing MWSs over those in the low-frequency bin, whereas L2 speakers did ($b = 0.293$, 95% CI = [0.097, 0.489], $p = .029$). In addition, L1 speakers showed significantly or marginally significantly greater advantages in processing MWSs in the very-high- ($b = 0.407$, 95% CI = [0.099, 0.714], $p = .041$) and high-frequency bins ($b = 0.222$, 95% CI = [0.027, 0.417], $p = .080$) relative to the medium-frequency bin, whereas L2 speakers did not show such significant differences. Taken together, such results suggest that the processing advantage of MWSs seems to follow an increasing pattern as frequency increases (see Figure 1).

Regarding association strength, L1 speakers' processing advantage of MWSs across association strength bins ranged from 0.481 to 0.760 for MI, from 0.320 to 0.686 for forward delta *P*, and from 0.438 to 0.662 for backward delta *P*. For L2 speakers, the

processing advantage of MWSs ranged from 0.160 to 0.648 for MI, from 0.268 to 0.669 for forward delta *P*, and from 0.253 to 0.618 for backward delta *P*. Subsequent multiple comparisons revealed that both L1 speakers and L2 speakers showed significantly or marginally significantly greater processing advantages for MWSs in the medium MI bin than in the low- (L1 speakers: $b = 0.260$, 95% CI = [0.126, 0.394], $p = .025$; L2 speakers: $b = 0.335$. 95% CI = [0.111, 0.559], $p = .033$) and very-high MI bins (L1 speakers: $b = 0.279$, 95% CI = [0.034, 0.524], $p = .052$; L2 speakers: $b = 0.487$. 95% CI = [0.181, 0.794], $p = .009$). Moreover, L1 speakers and L2 speakers' processing advantages of MWSs in the medium- (L1 speakers: $b = 0.364$, 95% CI = [0.180, 0.547], $p = .027$; L2 speakers: $b = 0.396$, 95% CI = [0.111, 0.681], $p = .038$) and high-forward-delta-*P* (L1 speakers: $b = 0.366$, 95% CI = [0.111, 0.621], $p = .019$; L2 speakers: $b = 0.239$, 95% CI = [0.015, 0.463], $p = .088$) bins were significantly or marginally significantly greater than those in the low-forward-delta-*P* bin.

However, there were notable differences in association measures between L1 speakers and L2 speakers. First, the Wald tests revealed marginally significant main effects of MI ($F = 4.98$, $p = .051$), forward delta *P* ($F = 4.18$, $p = .073$), and backward delta *P* ($F = 3.17$, $p = .096$) for L1 speakers but not for L2 speakers. Second, L2 speakers showed a smaller processing advantage for MWSs in the very-high MI bin relative to in the medium- and high MI bins (medium: $b = -0.487$, 95% CI = [-0.794, -0.181]; high: $p = .009$) ($b = -0.271$, 95% CI = [-0.544, -0.002], $p = .075$), whereas, L1 speakers did not exhibit such differences. Third, the multiple comparisons indicated that L2 speakers' processing advantage for MWSs in the very-high forward delta *P* bin was marginally significantly smaller than that in the medium bin ($b = -0.401$, 95% CI = [-0.749, -0.053], $p = .053$). In contrast, L1 speakers showed a significantly greater processing advantage for MWSs in the very-high forward delta *P* bin than in the low bin ($b = 0.359$, 95% CI = [0.118, 0.599], $p = .015$). Last, L1 speakers' processing advantage for MWSs in the medium ($b = 0.225$, 95% CI = [0.081, 0.368], $p = .046$) backward delta *P* bins was significantly greater than that in the low bin, whereas L2 speakers showed only a marginally significantly greater processing advantage for MWSs in the medium bin relative to in the very-high backward delta *P* bin ($b = 0.365$, 95% CI = [0.038, 0.692], $p = .054$). Analyses of the interactions between statistical regularities and language background revealed a significant difference between L1 speakers and L2 speakers in terms of processing advantages of MWSs. Specifically, this difference was observed in comparisons of the processing advantages between the low- and medium-frequency bin ($b = 0.232$, 95% CI = [0.081, 0.382], $p = .012$) and the low and very-high forward delta *P* bin ($b = -0.173$, 95% CI = [-0.346, 0.000], $p = .088$, see Appendix S15).

For both L1 speakers and L2 speakers, collocations (L1 speakers: Hedges's $g = 0.629$, 95% CI = [0.327, 0.930], $p = .001$; L2 speakers: Hedges's $g = 0.530$, 95% CI = [0.199, 0.861], $p = .013$), lexical bundles (L1 speakers: Hedges's $g = 0.258$, 95% CI = [0.138, 0.378], $p = .025$; L2 speakers: Hedges's $g = 0.308$, 95% CI = [0.215, 0.401], $p = .001$), and phrasal verbs (L1 speakers: Hedges's $g = 3.309$, 95% CI = [3.151, 3.468], $p = .016$; L2 speakers: Hedges's $g = 0.605$, 95% CI = [0.520, 0.691], $p = .046$) demonstrated significant processing advantages. Moderator analyses of MWS type indicated that neither L1 speakers nor L2 speakers had a significant processing advantage for binominals (Figure 2). Subsequent multiple comparisons revealed that both L1 speakers and L2 speakers had a larger processing advantage for phrasal verbs than for idioms (L1 speakers: $b = 2.927$, 95% CI = [2.655, 3.199], $p < .001$; L2 speakers: $b = 0.532$, 95% CI = [0.395, 0.670], $p < .001$) and lexical bundles (L1 speakers: $b = 3.051$, 95% CI =[2.853, 3.250], $p < .001$; L2 speakers: $b = 0.298$, 95% CI =[0.171, 0.424], $p = .004$).
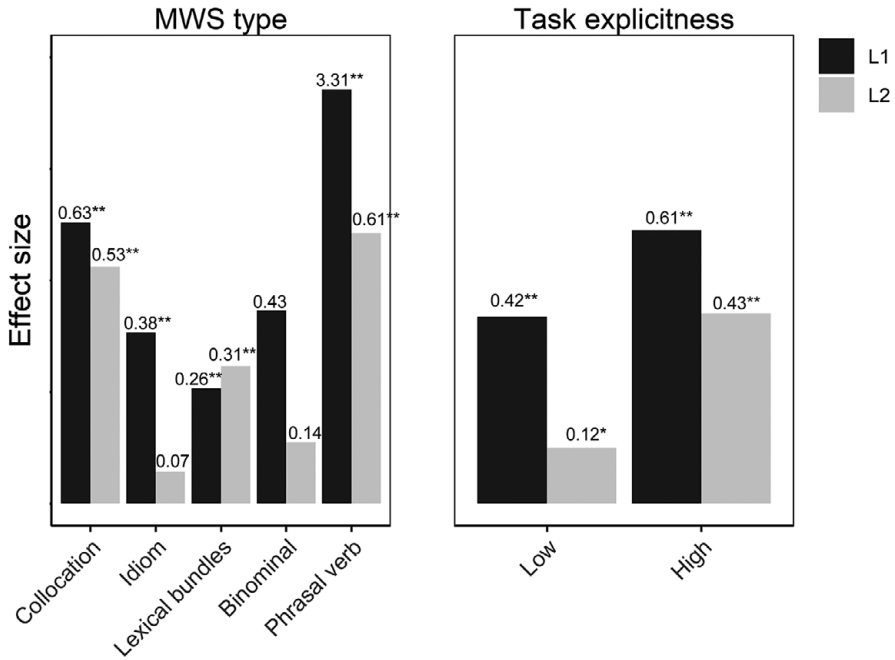
Figure 2. . Processing advantages of MWSs broken down by MWS type, task explicitness and language background. *p < .10, **p < .05.

Unlike L1 speakers (idiom: Hedges's g = 0.383, 95% CI = [0.161, 0.604], p = .007), L2 speakers did not demonstrate a processing advantage for idioms, and L2 speakers' processing advantages of collocations (b = 0.457, 95% CI = [0.109, 0.805], p = .022) and lexical bundles (b = 0.235, 95% CI = [0.092, 0.377], p = .008) were significantly greater than that of idioms. Furthermore, L1 speakers had a greater processing advantage for phrasal verbs than for collocations and binominals (collocation: b = 2.681, 95% CI = [2.341, 3.021], p = .001; binominal: b = 2.877, 95% CI = [2.481, 3.274], p < .001). Additionally, L1 speakers had a smaller (marginally significant) processing advantage for lexical bundles than for collocations (b = -0.370, 95% CI = [-0.695, -0.046], p = .077). Such differences were not significant in L2 speakers. Our analyses of the interaction between MWS type and language background revealed that L2 speakers exhibited more pronounced disadvantages in processing phrasal verbs (b = -2.612, 95% CI = [-2.737, -2.488], p < .001) relative to collocations than did L1 speakers. In contrast, L1 speakers experienced a marginally significantly larger disadvantage in processing lexical bundles relative to collocations than did L2 speakers (b = -0.123, 95% CI = [-0.237, -0.010], p = .084).

Last, regarding the moderating effect of task explicitness, both L1 speakers and L2 speakers showed significant or marginally significant processing advantages of MWSs for low- (L1 speakers: Hedges's g = 0.419, 95% CI = [0.184, 0.653], p = .004; L2 speakers: Hedges's g = 0.124, 95% CI =[0.017, 0.230], p = .054) and high-explicitness tasks (L1 speakers: Hedges's g = 0.612, 95% CI = [0.328, 0.895], p < .001; L2 speakers: Hedges's g = 0.426, 95% CI = [0.226, 0.625], p < .001). However, there was no moderating effect of task explicitness on the processing advantage of MWSs in L1 speakers (see Table 3), whereas the processing advantage of MWSs in L2 speakers

significantly increased when experimental tasks were more explicit ($b = 0.302$, 95% CI = [0.076, 0.528], $p = .020$; see Table 4). The analysis examining the interaction between task explicitness and language background did not yield significant results.

Discussion

This study sought to measure the processing advantage of MWSs in comparison with freely combined word combinations. Additionally, we explored whether this advantage is influenced by statistical regularities, types of MWSs, and task explicitness. In the following section, we will summarize and discuss the findings that address these research questions, beginning by examining similar patterns identified for both L1 speakers and L2 speakers and then proceeding to highlight differences observed between them.

The Processing advantage of MWSs

Overall, results obtained in this meta-analysis support the assertion that MWSs—despite being a complex and multifaceted construct (Biber, 2009)—offer an advantage over novel word combinations. The robustness of the processing advantage of MWSs also consolidates our belief that conventionalized multiword patterns can reduce our cognitive effort (Wray & Perkins, 2000) and help alleviate the time pressure placed on language users (Christiansen & Chater, 2016). Furthermore, our results suggest that various types of MWSs are represented in the memory and are the essential building blocks of language. Theoretically, our findings are in line with usage-based approaches (e.g., Christiansen & Arnon, 2017; Goldberg, 1995), which view language as an inventory of symbolic units of various sizes shaped by linguistic experience and predict that larger-than-word units are represented in the mental lexicon and share common cognitive mechanisms with single words.

   In terms of the magnitude of effect sizes, Plonsky and Oswald (2014) provided benchmarks for within-subject studies, indicating that an effect size of 0.6 is small, 1.0 is medium, and 1.4 is large. However, it should be noted that the benchmarks provided by Plonsky and Oswald (2014) are suitable for effect sizes in L2 research when accuracy is used as the outcome variable, which may not be as applicable when dealing with RT measures (Avery & Marsden, 2019). Given that there have been relatively few studies reporting effect sizes calculated from RTs using Cohen's *d* family of effect sizes, it would be difficult to conclude that the processing advantage of MWSs found in our meta-analysis (overall: Hedges's $g = 0.417$; 95% CI = [0.276, 0.558], $p < .001$; L1 speakers: Hedges's $g = 0.475$, 95% CI = [0.313, 0.636], $p < .001$; for L2 speakers: Hedges's $g = 0.348$, 95% CI = [0.220, 0.475], $p < .001$) is small. After all, the relatively small effect size calculated by RTs may be due to the nature of the data because RTs have larger standard deviations relative to their means, which can decrease effect sizes calculated using Cohen's *d* (Avery & Marsden, 2019; Brysbaert & Stevens, 2018).

Influences of moderators on the processing advantage of MWSs
Statistical regularities
Our meta-analysis showed that the processing advantages of MWSs were small to medium and significant across different bins of phrasal frequency and association strength for both L1 speakers and L2 speakers. Many studies in the literature assume

that processing advantage is restricted to word strings that are highly frequent or strongly associated. Methodologically, they usually have adopted a corpus-based approach and operationalized MWSs as word combinations that meet certain criteria of phrasal frequency (Biber et al., 1999; Wolter & Gyllstad, 2013) or association strength (e.g., Yi, 2018; Yi et al., 2017). However, our results seem to suggest that processing advantage applies to word strings across the whole continuum of statistical regularities. Consequently, it may not be a sound practice to limit the selection of MWSs to certain statistical thresholds (after all, they are arbitrary). Instead, it could be advantageous to switch to a continuous approach (Arnon & Snider, 2010) and extract MWSs from a wider range.

We also found that statistical regularities, including both phrasal frequency and association strength, moderate the processing advantage of MWSs. For both L1 speakers and L2 speakers, the processing advantage of MWSs seems to increase gradually as word combinations become more frequent. Interestingly, the processing advantage of MWSs seems to follow an inverted U-shaped curve (especially for L2 speakers), which peaks when constituent words within MWSs are not associated extremely strongly or extremely loosely (see Figure 1). Such moderating patterns indicate that both L1 speakers and L2 speakers are sensitive to statistical regularities when processing MWSs. Moreover, L1 speakers and L2 speakers are likely to share common statistical mechanisms when processing MWSs in real time (Yi, 2018). Regarding the different moderating patterns of phrasal frequency and association strength, they might result from the potential interaction between these two types of statistical information. Previous research has reported that higher frequency counts are associated with faster RTs when processing MWSs, which fits in well with the linear increasing pattern of the processing advantage of MWSs observed in this meta-analysis. Many studies (e.g., Ellis et al., 2008; Öksüz et al., 2021) have found that MWSs that are more strongly associated tend to be processed faster than those that are less strongly associated, at least for L1 speakers. However, an opposite pattern was also found in Yi (2018) for both L1 speakers and L2 speakers. Such contrasting results might relate to the difference in the frequency range of MWSs. Compared with Öksüz et al. (2021), Yi (2018) selected MWSs from a wider frequency range, with some word combinations occurring much less frequently in the corpus. Given that extremely strongly associated MWSs in this meta-analysis seem less frequent (see Appendix S16) and less familiar to participants (especially L2 speakers), it is possible that their processing speed could have been slowed down, leading to a reduced processing advantage compared with less strongly associated word sequences. However, such an explanation is tentative, and more studies will be needed to validate it.

MWS Type

Regarding the moderating effect of MWS type, first, despite that not all subtypes of MWSs enjoy certain processing advantages over novel word combinations, both L1 speakers and L2 speakers process collocations, lexical bundles, and phrasal verbs significantly faster than they do multiword units that are generated creatively. Second, we didn't find the processing advantage of binomials for both L1 speakers and L2 speakers. Binomials are typically semantically transparent and consist of associated words arranged in a preferred order. Sonbul et al. (2023) discovered that nonnative speakers tend to disregard the conventional word order of both familiar (e.g., black and white) and newly encountered binomials (e.g., bags and coats) in their L2. Conversely, Conklin and Carrol (2021) discovered that native speakers exhibited sensitivity to both

established (e.g., time and money) and novel binomials (e.g., wires and pipes) in their L1. Based on such findings, we posit that language users might not be sensitive to the order of binomials when encountered in a second language, especially during the early stage of lexical acquisition. However, it is noteworthy that binomials demonstrated a marginally significant processing advantage when considering L1-L2 mixture data (see Table S13.1 in Appendix S13). Therefore, the lack of a processing advantage for binomials among both L1 and L2 speakers could also potentially be attributed to the relatively limited number of effect sizes available. Third, we also found that the processing advantage of phrasal verbs (e.g., take off) is significantly greater than that of idioms and lexical bundles for both L1 speakers and L2 speakers. Nevertheless, given that only one study (i.e., Kim & Kim, 2012) on L1 speakers and L2 speakers' processing of phrasal verbs was included, future studies will be needed to check the magnitude of the processing advantage of phrasal verbs.

Task explicitness
Motivated by recent studies on the measurement of implicit and explicit knowledge using various tasks (e.g., Suzuki & DeKeyser, 2015, 2017), we operationalized task explicitness as the degree to which explicit knowledge is involved. We coded experimental tasks based on the presence of time pressure, availability of contextual support, and involvement of metalinguistic knowledge during the processing of MWSs. As an exploratory attempt, we found that both L1 speakers and L2 speakers exhibited significant processing advantages for MWSs over novel word combinations regardless of the explicitness of experimental tasks. Yi (2018) suggested that L1 speakers might process MWSs implicitly, whereas L2 speakers might process MWSs explicitly. According to Yi (2018), one might predict that L2 speakers would not show a processing advantage for MWSs in implicit tasks, which was not observed in our meta-analysis. Our findings highlight the robustness of the processing advantage for MWSs across both language groups and suggest that L2 speakers may also engage in implicit processing of MWSs under certain conditions (such as natural reading).

Differences between L1 speakers and L2 speakers

As mentioned in the previous section, L1 speakers and L2 speakers share similarities in processing MWSs. However, our meta-analysis also shows that they differ in the following aspects when processing MWSs in real-time.

First, L1 speakers' processing advantage for MWSs is significantly greater than that of L2 speakers (see Appendix S13). For both L1 speakers and L2 speakers, the overall effect size was determined by averaging across various types of MWSs that differed in statistical regularities and task explicitness. Given that L1 speakers exhibited more robust processing-advantage patterns than L2 speakers across all levels of the moderators, it is unsurprising that the overall processing advantage of MWSs for L1 speakers is greater than that for L2 speakers.

Our meta-analysis found that both L1 speakers and L2 speakers exhibited significant processing advantages for MWSs across the continuum of frequency and association strength. Moreover, their processing advantages of MWSs followed similar changing patterns (see Figure 1) as multiword units became more frequent and associated. Nevertheless, we also found differences in the moderating patterns of statistical regularities on the processing advantage of MWSs, with some between-bin comparisons being significant or nonsignificant for either L1 speakers or L2 speakers. From a

usage-based perspective, such differences are most likely to be attributed to the differences in the language experience of L1 speakers and L2 speakers. Another notable difference between L1 speakers and L2 speakers lies in their sensitivity to the direction of the association between constituent words within multiword units. As mentioned earlier, when backward delta *P* was adopted as the measure of association strength, only L1 speakers demonstrated a significant facilitating effect of increased backward delta *P* on the processing advantage of MWSs. Given that backward delta *P* measures the predictability of the last word following other words in a sequence, such a result suggests that L2 speakers—unlike L1 speakers—might not be sensitive to the co-occurring probability of MWSs in a backward fashion.

Significant differences between L1 speakers and L2 speakers with respect to the moderating effect of MWS type on the processing advantage for MWSs were also found. Specifically, the small-to-medium processing advantage of idioms as found in L1 speakers was not replicated in L2 speakers. The discrepancy in the processing advantage of idioms between L1 speakers and L2 speakers might result from the distinct approaches taken by these two groups of language users. According to the dual-route model (van Lancker Sidtis, 2012), L1 speakers process idioms via either direct retrieval or computational analysis, with the former being the default route for lexicalized idiomatic units and the latter responsible for novel phrases. In comparison, the literal saliency model (Cieślicka, 2006) proposes that L2 speakers tend to automatically decompose idioms and rely on literal analysis of constituent words to compute the meaning of such units. As a result, unlike L1 speakers, L2 speakers might not be able to establish fast form-meaning mappings. Instead, they might have to take an additional step to reject and suppress the literal interpretation of idioms, which leads to no processing advantage over novel word combinations. Despite these theoretical accounts, the absence of processing advantages of idioms in L2 speakers might result from other sources as well. For instance, Shi et al. (2022) found that more proficient L2 speakers showed less delay when processing figurative collocations than less proficient L2 speakers did. They proposed that L2 speakers may shift from the computation approach to the direct retrieval approach once they become proficient enough in the target language. Taken together, future studies are needed to explore whether processing advantages of idioms will show up when L2 speakers with native-like proficiency are recruited.

We also found that L2 speakers' processing advantages of collocations and lexical bundles are significantly greater than that of idioms, which were not found in L1 speakers. Additionally, our findings suggest that the processing advantage of collocations is trending toward being larger than that of lexical bundles in L1 speakers. Our interaction analyses further showed that the processing advantage of collocations relative to lexical bundles was marginally significantly greater for L1 speakers than for L2 speakers. Collocations are structurally self-contained and semantically compositional, although their constituent words can be interpreted either literally (e.g., *break the vase*) or figuratively (e.g., *break the rule*). Lexical bundles (e.g., *is one of the*) are semantically compositional but structurally incomplete, which usually crosses syntactic boundaries. In contrast, idioms are not semantically compositional and are usually interpreted figuratively. Given that lexical bundles are structurally less complete than collocations, the above results suggest that L1 speakers might be more sensitive to the structural completeness of MWSs than L2 speakers. Furthermore, given that idioms differ from collocations and lexical bundles mainly in semantic figurativeness, L2 speakers' processing disadvantage of idioms relative to collocations may largely be

attributed to their lack of semantic representation of such multiword units as compared with L1 speakers.

Last, we found that the processing advantage of MWSs was not moderated by task explicitness for L1 speakers. In contrast, such a moderating effect was found for L2 speakers, with the processing advantage of MWSs being greater when experimental tasks are more explicit. Given that the interaction between task explicitness and language background was nonsignificant, we interpret such results as evidence supporting a trend in the discrepancy between L1 speakers' and L2 speakers' sensitivity to task explicitness. Two possible reasons might explain such a discrepancy. First, due to limited exposure to the target language under naturalistic conditions, L2 speakers' knowledge of MWSs might be more explicit than that of L1 speakers. Another possibility may lie in the distinct approach taken by L1 speakers and L2 speakers when processing MWSs. As already mentioned, L2 speakers tend to analyze the input into individual words (Cieślicka, 2006; Wray, 2012) and process MWSs using a computational approach. In contrast, L1 speakers can directly retrieve MWSs based on underlying statistical regularities, which are tallied implicitly (Ellis, 2012). Evidence in favor of this possibility can also be found in Yi (2018), in which L1 speakers' response time was facilitated by implicit aptitude when judging whether word combinations are acceptable under time pressure. More studies will be needed to further examine the distinct moderating role of task explicitness on the processing of MWSs for L1 speakers and L2 speakers.

Limitations and future studies

This meta-analysis has some limitations that should be considered. First, our analysis only included studies that used perception tasks and did not consider production tasks. Therefore, future studies could incorporate both types of tasks and compare the processing advantage of MWSs between them. Second, to ensure consistency in the definition of MWSs and comparability of results, our meta-analysis only included studies that used English stimuli. Future studies replicating our meta-analysis with languages other than English are needed. Third, we only incorporated one study that focused on the processing of phrasal verbs (Kim & Kim, 2012). To validate our findings regarding the processing advantage of phrasal verbs, more studies are needed. Fourth, as our meta-analysis is the first to code experimental tasks based on their degree of explicit knowledge required, further research is necessary to extend our findings. Finally, given the scope of this meta-analysis, we only examined the moderating effects of statistical regularities, MWS type, task explicitness, and language background on the processing advantage of MWSs. Further studies could explore other factors that might influence the processing of multiword units, such as language aptitude (Yi, 2018).

Conclusion

This meta-analysis synthesized empirical research on the processing advantage of MWSs during the past 2 decades. Our results confirm that MWSs enjoy a processing advantage over novel word combinations during online tasks and that such a processing advantage is moderated by statistical regularities (i.e., phrasal frequency, association strength), MWS type, and task explicitness. Furthermore, L1 speakers and L2 speakers show both commonalities and differences when processing MWSs in real

time, which suggests a need to treat L1 speakers and L2 speakers as nonhomogenous groups in future studies.

The findings in our meta-analysis have important theoretical implications. Specifically, the robustness of the processing advantage of MWSs (at least for L1 speakers) indicates that larger-than-word units are lexically represented in the mental lexicon and subject to statistical mechanisms. Such results lend support to the shift from the words-and-rules approach (Pinker & Ulman, 2002) to usage-based theories (Christiansen & Arnon, 2017; Goldberg, 1995) given that the former views MWSs (except idioms) as phrases that are generated based on grammatical rules and predict no processing advantage for multiword units. Note, however, that the processing advantage of MWSs as confirmed in our meta-analysis should not be taken as evidence supporting that MWSs are necessarily processed as unanalyzed or holistic units (i.e., stored and retrieved whole from memory at the time of use; Wray & Perkins, 2000) because the incorporated studies were not designed to test the holistic representation issue (Siyanova-Chanturia, 2015) and analytic processing in parallel with holistic processing have been found even for fixed expressions such as idioms (e.g., Boulenger et al., 2012).

The processing advantage for various types of MWSs as revealed in this meta-analysis corroborates the viewpoint that MWSs are multifaceted and heterogeneous. Despite this, our results suggest that subtypes of MWSs share common cognitive mechanisms. They are processed significantly faster than novel word combinations and can compensate for the limitation in our cognitive resources. Methodologically, previous studies have defined MWSs based on different dimensions, following either the phraseological approach (Wray, 2000) or the corpus-based approach (Yi, 2018). The phraseological approach views multiword units as a continuum, with free combinations at one end and idioms at the other end. In contrast, the corpus-based approach defines MWSs as word strings that are highly frequent or strongly associated. Neither approach is fully supported by our results. For example, in our meta-analysis, the processing advantage of MWSs was found not only for idiosyncratic expressions but also for semantically transparent, less formulaic word combinations. Similarly, although the corpus-based approach predicts the processing advantage of highly frequent word strings such as lexical bundles, it does not predict the processing advantage of infrequent fixed expressions (e.g., not every idiom is highly frequent; Wray, 2012). Our results seem to bridge such methodological gaps and provide a common ground for these two distinct approaches by defining MWSs as cognitively privileged units that enjoy a processing advantage over word combinations generated creatively.

References

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

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

Avery, N., & Marsden, E. (2019). A meta-analysis of sensitivity to grammatical information during self-paced reading: Towards a framework of reference for reading time effect sizes. *Studies in Second Language Acquisition*, *41*, 1055–1087. https://doi.org/10.1017/S0272263119000196

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*, 241–248. https://doi.org/10.1111/j.1467-9280.2008.02075.x

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, *14*, 275–311. https://doi.org/10.1075/ijcl.14.3.08bib

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

Borenstein, M., Cooper, H., Hedges, L., & Valentine, J. (2009). Effect sizes for continuous data. In H. Copper, L. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-Analysis* (pp.221–236). Russell Sage Foundation.

Boulenger, V., Shtyrov, Y., & Pulvermüller, F. (2012). When do you grasp the idea? Meg evidence for instantaneous idiom understanding. *Neuroimage*, *59*, 3502–3513. https://doi.org/10.1016/j.neuroimage.2011.11.011

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*, Article 9. https://doi:10.5334/joc.10

Cappelle, B., Shtyrov, Y., & Pulvermüller, F. (2010). Heating up or cooling up the brain? MEG evidence that phrasal verbs are lexical units. *Brain and Language*, *115*, 189–201.

Carrol, G., & Conklin, K. (2014). Getting your wires crossed: Evidence for fast processing of L1 idioms in an L2. *Bilingualism: Language and Cognition*, *17*, 784–797. https://doi.org/10.1017/S1366728913000795

Carrol, G., & Conklin, K. (2020). Is all formulaic language created equal? Unpacking the processing advantage for different types of formulaic sequences. *Language and Speech*, *63*, 95–122. https://doi.org/10.1177/0023830918823230

Christiansen, M. H., & Arnon, I. (2017). More than words: The role of multiword sequences in language learning and use. *Topics in Cognitive Science*, *9*, 542–551. https://doi.org/10.1111/tops.12274

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, Article e62. https://doi.org/10.1017/S0140525X1500031x

Cieślicka, A. (2006). Literal salience in on-line processing of idiomatic expressions by second language learners. *Second Language Research*, *22*, 115–144. https://doi.org/10.1191/0267658306sr263oa

Columbus, G., & Wood, D. (2010). Processing MWUs: Are MWU subtypes psycholinguistically real? In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 194–212). Continuum.

Conklin, K., & Carrol, G. (2021). Words go together like "bread and butter": The rapid, automatic acquisition of lexical patterns. *Applied Linguistics*, *42*, 492–513. https://doi.org/10.1093/applin/amaa034

Durrant, P. L. (2008). *High frequency collocations and second language learning* [*Unpublished doctoral dissertation*]. University of Nottingham, Nottingham, UK.

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

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172. https://doi.org/10.1017/s0272263105050096

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17–44. https://doi.org/10.1017/S0267190512000025

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, *42*(3), 375–396. https://doi.org/10.2307/40264474

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

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *p* values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*, 337–350. https://doi.org/10.1007/s10654-016-0149-3

Gries, S. T. (2013). 50-something years of work on collocations what is or should be next. *International Journal of Corpus Linguistics*, *18*, 137–165. https://doi.org/10.1075/ijcl.18.1.09gri

Gries, S. T. (2022). What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies*, *5*, 1–33. https://doi.org/10.1075/jsls.21028.gri

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

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

Hanna, J., Cappelle, B., & Pulvermüller, F. (2017). Spread the word: MMN brain response reveals whole-form access of discontinuous particle verbs. *Brain and Language*, *175*, 86–98. https://doi.org/10.1016/j.bandl.2017.10.002

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*, 39–65.

Jeong, H., & Jiang, N. (2019). Representation and processing of lexical bundles: Evidence from word monitoring. *System*, *80*, 188–198. https://doi.org/10.1016/j.system.2018.11.009

Jiang, N. (2012). *Conducting reaction time research in second language studies*. Routledge.

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *Modern Language Journal*, *91*, 433–445. https://doi.org/10.1111/j.1540-4781.2007.00589.x

Kim, S. H., & Kim, J. H. (2012). Frequency effects in L2 multiword unit processing: Evidence from self-paced reading. *TESOL Quarterly*, *46*, 831–841. https://doi.org/10.1002/tesq.66

Kyriacou, M., Conklin, K., & Thompson, D. (2021). When the idiom advantage comes up short: Eye-tracking canonical and modified idioms. *Frontiers in Psychology*, *12*, 675046. https://doi.org/10.3389/fpsyg.2021.675046

Lindstromberg, S., & Eyckmans, J. (2020). The effect of frequency on learners' ability to recall the forms of deliberately learned L2 multiword expressions. *International Journal of Applied Linguistics*, *171*, 2–33. https://doi.org/10.1075/itl.18005.lin

Liversedge, S. P., Paterson, K. B., & Pickering, M. J. (1998). Eye movements and measures of reading time. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 55–75). Elsevier Science Ltd.

Long, M. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, R. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39–52). John Benjamins.

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

Pellicer-Sánchez, A., Siyanova-Chanturia, A., & Parente, F. (2022). The effect of frequency of exposure on the processing and learning of collocations: A comparison of first and second language readers' eye movements. *Applied Psycholinguistics*, *43*, 727–756.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*, 456–463.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.

Pustejovsky, J. E. (2015). Measurement-comparable effect sizes for single-case studies of free-operant behavior. *Psychological Methods*, *20*, 342–359. https://doi.org/10.1037/met0000019

Pustejovsky, J. E. (2022). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections* (R package version 0.5.5) [Computer software]. https://CRAN.R-project.org/package=clubSandwich

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. URL https://www.R-project.org/

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, *12*, 110–114. https://doi.org/10.1207/S15327078IN0402_07

Shi, J., Peng, G., & Li, D. (2022). Figurativeness matters in the second language processing of collocations: Evidence from a self-paced reading experiment. *Language Learning*, *73*, 47–83. https://doi.org/10.1111/lang.12516

Siyanova-Chanturia, A. (2015). Collocation in beginner learner writing: A longitudinal study. *System*, *53*, 148–160. https://doi.org/10.1016/j.system.2015.07.003

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

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

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

Sonbul, S., El-Dakhs, D. A. S., Conklin, K., & Carrol, G. (2023). "Bread and butter" or "butter and bread"? Nonnatives' processing of novel lexical patterns in context. *Studies in Second Language Acquisition*, *45*, 370–392.

Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning*, *65*, 860–895. https://doi.org/10.1111/lang.12138

Suzuki, Y., & DeKeyser, R. (2017). The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning*, *67*, 747–790. https://doi.org/10.1111/lang.12241

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, *40*, 604–634. https://doi.org/10.3102/1076998615606099

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 151–173). Continuum.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, *61*, 569–613. https://doi.org/10.1111/j.1467-9922.2010.00622.x

van Lancker Sidtis, D. (2012). Two-track mind: Formulaic and novel language support a dual-process model. In M. Fraust (Ed.), *The handbook of the neuropsychology of language* (pp. 342–367). Blackwell.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48. https://doi.org/10.18637/jss.v036.i03

Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing. *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

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

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics*, *32*, 231–254. https://doi.org/10.1017/S026719051200013x

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication*, *20*, 1–28. https://doi.org/10.1016/S0271-5309(99)00015-

Yanagisawa, A., & Webb, S. (2021). To what extent does the involvement load hypothesis predict incidental L2 vocabulary learning? A meta-analysis. *Language Learning*, *71*, 487–536. https://doi.org/10.1111/lang.12444

Yanagisawa, A., Webb, S., & Uchihara, T. (2020). How do different forms of glossing contribute to L2 vocabulary learning from reading? *Studies in Second Language Acquisition*, *42*, 411–438. https://doi.org/10.1017/s0272263119000688

Yi, W. (2018). Statistical sensitivity, cognitive aptitudes, and processing of collocations. *Studies in Second Language Acquisition*, *40*, 831–856. https://doi.org/10.1017/s0272263118000141

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

Yi, W., Man, K., & Maie, R. (2023) Investigating L1 and L2 speaker intuitions of phrasal frequency and association strength of multiword sequences. *Language Learning*, *73*, 266–300. https://doi.org/10.1111/lang.12521