ITL - International Journal of Applied Linguistics

The effect of frequency on learners' ability to recall the forms of deliberately learned L2 multiword expressions

--Manuscript Draft--

Manuscript Number:	ITL-18005R3
Full Title:	The effect of frequency on learners' ability to recall the forms of deliberately learned L2 multiword expressions
Short Title:	The effect of frequency
Article Type:	Article
First Author:	Seth Lindstromberg, MPhil
Other Authors:	June Eyckmans, PhD
Corresponding Author:	Seth Lindstromberg, MPhil Hilderstone College Broadstairs, Kent UNITED KINGDOM
Funding Information:	
Section/Category:	Article
Keywords:	L2 multiword expressions, productive knowledge, deliberate learning, frequency effect, phonological forms, free recall
Abstract:	In incidental learning, vocabulary items with high or relatively high objective frequency in input are comparatively likely to be acquired. However, many single words and most multiword expressions (MWEs) occur infrequently in authentic input. It has therefore been argued that learners of school age or older can benefit from episodes of instructed or self-managed deliberate (or intentional) L2 vocabulary learning, especially when L2 is learned in an EFL environment and most especially when productive knowledge is the goal. A relevant question is whether the objective frequency of vocabulary items is an important factor in production-oriented deliberate L2 vocabulary learning. We report three small-scale interim meta-analyses addressing this question with regard to two-word English Adj-Noun and Noun-Noun expressions. The data derive from 8 original studies involving 406 learners and 139 different MWEs. Our results suggest that objective frequency has a weak, possibly negative effect in the deliberate learning of MWE forms.
Author Comments:	Dear Editor I have endeavored to make all the changes recommended by you and by the reviewer. Those changes are specified in a separate file following the figure. Thank you for finding such perspicacious reviewers. Sincerely Seth Lindstromberg
Order of Authors Secondary Information:	

The effect of frequency on learners' ability to recall the forms of deliberately learned L2 multiword expressions

For some years we have studied the acquisition of L2 multiword expressions (MWEs) in situations of deliberate learning. Frequently, reviewers have asked us to say more about the effect of frequency on the outcomes. But when looking at the data, we consistently failed to detect the strong positive frequency effect that some of our research colleagues assume must exist. A question suggested by this situation is: Why would a L2 researcher expect frequency to have an appreciable positive effect on learners' ability to produce targeted, L2 English MWEs in contexts of deliberate learning? Another question is: Does any such effect exist? We address these and certain related questions in later sections of this article. In doing so we use a number of terms which we should define without further ado.

Productive knowledge (or *active knowledge*) of lexical items consists in being able to say them or write them autonomously, which is a feat that learners tend to find even more difficult than dealing with the semantic components of vocabulary production (Nation, 2013). The best evidence of productive knowledge of a lexical item is success in producing the item in the absence of a 'retrieval cue' (e.g., Anderson, 2009) conveying information about phonological or orthographic form. Put another way, "The more one provides of a form of a word, the less one is testing productive knowledge" (Barcroft, 2015, p. 50). A strong test of productive knowledge is a free recall test. A cued recall test is a less strong test because, by definition, it furnishes form retrieval cues such as the first sound of a to-be-recalled word or

the first word of a to-be-recalled MWE. When we use the equivalent terms *retrieval* and *recall* further below, our focus is on learners' success in producing lexical *forms*.

Frequency can refer to the number of times that lexical items (especially lemmas and base forms) occur in a mega-corpus such as the Corpus of Contemporary American English, or COCA (Davies, 2008-2018).[1] Used this way the term frequency is about 'absolute frequency' (e.g., Crossley, et al., 2014) or, equivalently, 'objective frequency' (e.g., Meschyan & Hernandez, 2002). Researchers often use the objective frequency of this or that lexical item as a proxy for its 'experienced frequency'—which is, roughly, the number of times a given lexical item has actually been encountered and processed by a given (type of) individual. A person's experienced frequency of an item is presumed to be an important determinant of the quality of that item's representation in the person's mental lexicon (e.g., Perfetti & Hart, 2002). Unless noted otherwise, we will use the term *frequency* to refer to objective frequency as measured by frequency in COCA.

The term *incidental vocabulary learning* refers to vocabulary learning (or, equivalently, vocabulary acquisition) that comes as a by-product of message-focused reading or listening (Hulstijn, 2001). In contrast, *deliberate vocabulary learning* refers to vocabulary learning that occurs either when learners follow an instruction to try to learn stipulated vocabulary (with or without the expectation of being tested) or when learners engaged in self-managed study choose their own vocabulary to try to learn (cf., Barcroft, 2015, p. 51). Deliberate (or intentional) learning is likely to be associated with so-called 'explicit' or 'direct' teaching as described for example by Sonbul and Schmitt (2013). Accordingly, deliberate learning typically involves what has been called 'focus on forms', a type of engagement with vocabulary that is likely to result in richer, more elaborative mental processing of orthographic and phonetic forms than is usual in situations of incidental learning (Laufer, 2005). An important practical difference between situations of incidental

and deliberate learning is that in the latter it often happens that a learner encounters each targeted item the same predetermined number of times (Barcroft, pp. 51-52). For example, each targeted item may be included just once in a bilingual word list or set of vocabulary cards and in any associated set of test questions. Learners in such cases are generally able—albeit perhaps within a time limit—to decide how long or how often they will focus on any given item; even so in situations of deliberate learning there is unlikely to be any semblance of the huge variation in the number of item occurrences that is typical for incidental vocabulary learning. True, there have been a few studies in which, by design, learners were exposed to targeted items different, predetermined numbers of times—for instance, some items just once each and others four times each. In these studies the goal has been to investigate the effects of such variation (for reviews see Peters, 2014, 2016). However, our focus will *not* be on studies of this type.

Literature review

Frequency effects in general and in L1 in particular

Frequency effects have received considerable attention from researchers of language, language use and language learning. For example, a range of frequency effects have been documented in language change (Bybee, 2010), L1 acquisition (Ellis, 2002, Goldberg, 2006; Tomasello, 2003), and online processing of L1 vocabulary. Let us temporarily focus on the latter stream of research, in which frequency effects have been investigated for instance by means of lexical decision tasks for measuring the speed with participants recognize word forms (e.g., Gardner, Rothkopf, Lapan, & Lafferty, 1987; Glanzer & Ehrenreich, 1979; Gordon & Caramazza, 1982; Whaley, 1979) and picture naming tasks set in order to measure participants' speed in producing word forms (e.g., Bates, et al., 2003; Oldfield & Wingfield, 1968). In these investigations the key outcome data tend to be response latencies (i.e., time

lapses between cue and response); and the overwhelmingly typical result has been that latencies are comparatively short for relatively frequent words. Evidently, high frequency tends to facilitate processing—often strongly. One explanation for this is that a high frequency (HF) word affords more retrieval cues than a low frequency (LF) word does. That is to say, a HF word is comparatively likely to be stored in memory with connections to the meanings and forms of many other words. These connections are thought to make HF words comparatively retrievable because a thought or an experience that activates any part of a network of such connections has the potential also to activate the memory traces of a to-be-remembered item and so facilitate its recognition or recall (e.g., Diana & Reder, 2006; Storkel & Morrisette, 2002). One apparent example of such 'spreading activation' (Loftus & Loftus, 1974; Dell & Gordon, 2003) is the phenomenon known as 'phonological priming' (Goldinger, et al., 1992), whereby occurrence of a word in input has the potential to facilitate processing of a subsequently occurring word which shares one or more phonemes with the first word (e.g., *b*ull → *b*eer). We say more about phonological priming in a later section.

In a further stream of L1 research involving decontextualized vocabulary (generally L1 words or L1-like pseudowords) an early and amply replicated finding relates to experimental participants who are either asked to recall L1 words in any order from previously studied 'pure' lists comprising only HF words or only LF words. In subsequent *recall* tests participants do better on HF words encountered on pure HF lists than they do on pure LF lists (Hall, 1954). But in tests of *recognition* the advantage is reversed (e.g., McLeod & Kampe, 1996). One hypothesis regarding the LF advantage in recognition is that LF words are given more attention than HF words. A possible contributing factor here is the tendency for LF content words to be distinctive owing to systematic structural differences from HF content words. It has been found, for instance, that LF content words are more likely to include a nasal phoneme and less likely to include an alveolar (Pisoni, Nusbaum, Luce, &

Slowiaczek, 1985; for relevant discussion of the concept of distinctiveness see Dewhurst, Brandt, & Sharp, 2004). According to the 'distinctiveness hypothesis' a LF word is likely to be comparatively richly encoded during a study phase (on account of its formal distinctiveness) with the result that the mental representation of its form and/or meaning becomes more elaborate. It is also part of this hypothesis that the tendency for LF words to be formally distinctive means that they are comparatively likely to be encoded in association with 'episodic' information—that is, information about when and in what context the words were encountered—which is known to enhance likelihood of later recognition. A second, apparently complementary, hypothesis is that participants in an experiment—and perhaps also learners in a given lesson—are comparatively likely to perceive LF stimulus words as novel for the reason that these words will have been encountered prior to the experiment (or lesson) less often and in fewer contexts than HF stimulus words (Brown, 1976; Eysenck & Eysenck, 1980). This could matter not only because categorization of a word as novel entails noticing the word, but also because categorization of a word as novel is a precondition for and a possible trigger of word learning, the first phase of which is creation of a new mental representation for the word (Han, Storkel, & Yoshinaga-Itano, 2015). Such a representation is likely to remain activated for a while after its creation, and during this period of activation the word in question is comparatively recognizable if re-encountered. According to a third (also possibly complementary) hypothesis, LF words are given more attention than HF words and are relatively richly encoded because the processing of LF words requires extra time and additional mental resources (see, e.g., Dewhurst et al., 2004). (For reviews of the literature on frequency effects on lexical memory see Dewhurst et al.; Diana & Reder, 2006; and Criss, Aue, & Smith, 2011). We will return to these three hypotheses further below in connection with the processing and subsequent retrieval of MWEs.

In the previous paragraph we outlined the gist of findings regarding participants who studied *pure lists* of HF words or pure lists of LF words prior to a test of recall or recognition. But there have also been many studies in which participants have been tested on their ability to recall words in any order from so-called *mixed lists* of jumbled HF and LF targets. Results of these mixed-list studies may be more relevant to L2 learning than the pure list studies for the reason that on any given occasion learners are comparatively likely to encounter L2 words from a number of frequency bands. As it happens, though, these mixed-list studies have had disparate results. Lohnas and Kahana (2013), who reviewed the relevant literature (see also Dewhurst et al., 2004), attributed this state of affairs to the fact that previous researchers had not included *medium* frequency words in their lists. When Lohnas and Kahana tested 132 participants on their ability to recall L1 words (N = 894) studied in mixed lists that did include medium frequency words, they found a U-shaped pattern, such that HF and LF words were both recalled better than ones of middle frequency. These researchers suggested that in any-order, mixed-list trials the retrieval of HF words and the retrieval of LF words are facilitated by *different* frequency effects. That is, for HF words a key facilitating factor may be a greater number of neural connections between the mental representations of these words and other known words whereas for LF words a facilitating factor may be the greater likelihood of LF words being richly encoded during the study phase on account of their distinctiveness (cf., DeLosh & McDaniel, 1996). An additional facilitating factor may be that people expect LF words to be comparatively hard to remember. Therefore, *especially when they are learning deliberately,* they devote extra processing time and resources to these words (Watkins, LeCompte, & Kim, 2000). A third facilitating factor may be a novelty effect as outlined not far above. As yet there is no uncontroversial model that accounts for all the pertinent effects seen in recognition and in recall of lexical items (for discussion see Criss et al., 2011). Further below we refer again to factors that may facilitate the recall of LF items.

Effects of frequency in incidental learning of L1 and L2 vocabulary

Most L1 vocabulary acquisition occurs during incidental exposure to input, with frequency being a strong facilitating factor (Jenkins, Stein, & Wysocki, 1984). Thus, the more frequent a L1 vocabulary item is, the lower the age at which it is likely to be acquired (e.g., Stadthagen-Gonzales & Davis, 2006; Tellings, Coppens, Gelissen, & Schreuder, 2013), which is in line with evidence that frequency is a positive factor in the incidental acquisition of *L2* vocabulary (Ellis, 2002; Nation, 2013; Webb, Newton, & Chang, 2013). Constraining factors are, first, that L2 learners tend to have far fewer opportunities to learn vocabulary incidentally than native speakers (NSs) do and, second, that the vast majority of words in a natural language occur infrequently, as per Zipf's law. To illustrate, of the 600,000 non-obsolete headwords in the *Oxford English Dictionary* only 5.2% occur more often than once per million words and only 1.2% occur more often than ten times per million words (Oxford University Press, 2017). Frequencies of MWEs tend to be even lower (Moon, 1998), which is what one might expect given that no MWE can have a frequency greater than any of its constituent words. Accordingly, growth of productive L2 vocabulary knowledge through incidental exposure tends to proceed slowly, especially with respect to MWEs (for reviews see Henriksen, 2013; Laufer & Waldman, 2011; Peters, 2016); and learners in an EFL environment are likely to develop *productive* knowledge of L2 vocabulary in general at a particularly slow rate (e.g., Laufer, 2005; Nation, 2013; Schmitt, 2008; Waring & Takaki, 2003). Because L2 vocabulary can be learned more efficiently in situations of deliberate learning than in situations of incidental learning (e.g., Laufer, 2005; Nation, 2013; Schmitt, 2008), it has been argued that a well-rounded plan for learning L2 vocabulary should include a good amount of instructed and self-managed deliberate learning, particularly when learning takes place in an EFL environment (Nation, 2013).

Deliberate learning may be especially beneficial in the case of MWEs owing to the fact that they are comparatively liable to be overlooked in situations of incidental learning (e.g., Boers & Lindstromberg, 2009, Ch. 3; Boers, Lindstromberg & Eyckmans, 2014a). One reason MWEs may be more susceptible to being overlooked is that a substantial proportion of MWEs are made up of comparatively frequent words (Martinez & Murphy, 2011; Nation, 2013) which learners tend not to focus on. To elaborate, it has been found that individuals engaged in message-focused reading spend less time looking at HF words than at LF words, a pattern of behavior that is especially evident among readers in L2 (Cop, et al., 2015). Another impediment to incidental learning of L2 MWEs is indicated by evidence that many learners—including quite proficient ones—find it hard to identify the MWEs that they encounter in written input (Martinez & Murphy, 2011) even in situations where they have been instructed to mark all the MWEs they see and have been given ample time to do so (Eyckmans, Boers, & Stengers, 2007).

Frequency effects on ability to recall deliberately learned L2 vocabulary

There are reasons to doubt that everything that is true about effects of frequency on a person's ability to process L1 vocabulary is true also with respect to vocabulary in L2. For one thing, effects of frequency in the processing of L1 vocabulary seem to be appreciably moderated by the subjective familiarity of the vocabulary as well as by the age at which it was acquired. It is likely though that neither of these two variables matters as much for L2 learners as for same-age native-speakers (NSs) since many L2 learners will have acquired few if any L2 words at an early age and will have had less time and fewer opportunities to develop deep familiarity with a wide range of L2 vocabulary items. Accordingly, possible differences between effects of frequency in the processing of L1 and L2 vocabulary have been the object of a number of investigations. Most relevant to our purposes are ones relating to MWEs (for overviews see Siyanova-Chanturia & Janssen, 2018; Sonbul & Schmitt, 2013.)

However, frequency effects in *deliberate* learning seem to have received comparatively little attention in this stream of research; and little more can be said about the available results than that they suggest it would be unsafe to assume that frequency effects in deliberate L2 vocabulary learning will mirror frequency effects seen in L1 and L2 incidental learning or in L1 online processing (see Peters, 2014 & 2016, and Sonbul & Schmitt, 2013, for discussion of some relevant issues). For example, Lotto & de Groot (1998) found that frequency is associated with easier deliberate learning of *single words*. But there is also evidence that in deliberate word learning frequency is less important than it is in incidental word learning (Hamrick & Rebuschat, 2014) and that frequency matters much less in deliberate word learning than do concreteness of word meaning and cognate status (i.e., whether a word has a L1 cognate) (de Groot & Keijzer, 2000). A meta-analysis of 19 primary studies reported by Durrant (2014) found mostly small or modest positive effects of frequency on learners' knowledge of collocations following explicit instruction. However, in those studies there was little or no focus on free recall of whole MWEs. For example, 63% of the posttests were multiple choice tests, meaning that they had to do with form recognition rather than with form recall, which may explain the predominance of observed positive effects.

As might be expected, investigation of frequency effects in MWE learning are relatively complex because more measures of frequency may need to be considered than is the case for investigations of effects of frequency in word learning and because the measures may not strongly correlate, as will be seen. One measure of frequency is of course the frequency of each MWE as a whole. A second measure is the combined frequency of the constituent content words or, equivalently for statistical analysis, their average frequency. A third but so far little used measure is the frequency of the least frequent constituent content word. A reason for considering this last measure is evidence referred to further above that LF words are particularly likely to attract learners' attention whereby the forms of LF words tend

to be better remembered than the forms of HF words, all else being equal. Further, in the context of cued recall it may make sense to separately consider the frequencies of the words used as cues and the words figuring as responses (e.g., *green*$_{\text{.Cue}}$→*grass.*$_{\text{Response}}$). Additional frequency measures have been used as well (e.g., bigram and trigram frequencies) but they are beyond the scope of this study (see, for instance, Durrant, 2014).

Preliminary observations that led us to question the importance of frequency in deliberate vocabulary learning were as follows. In one experimental study of the deliberate learning of potentially *novel* L2 English figurative idioms, Eyckmans and Lindstromberg (2016) calculated first order correlations and values of r^2 with respect to pretest-to-delayed-posttest gain scores and the frequencies of (a) whole idioms, (b) all constituent content words, and (c) the content words figuring as responses in the tests of cued recall. In that study we found that no value of r^2 exceeded .009. Moreover, the first order Pearson's correlations were not all signed in the same direction. These results suggest that frequency effects could be negligible. However, in that study, and in others involving *novel* MWEs, learners had to cope not only with unfamiliar forms but also with the requirement to learn new form-meaning mappings and perhaps even new meanings. This is hardly a situation in which it would be easy to isolate the effect of frequency on productive form knowledge. More relevant to the study we report below is a quasi-experimental study of immediate *cued* recall of studied L2 English MWEs that are likely to have been at least somewhat familiar to participating learners. In this study Lindstromberg and Eyckmans used a multiple linear regression model that included the following explanatory variables: (a) presence-of-assonance, (b) concreteness-imageability, (c) mutual information (MI) score, (d) whole-MWE-frequency, (e) frequency of first word, and (f) frequency of second word. We found that the three frequency variables accounted for less than 1% of variation in test scores; moreover, the first order Pearson's correlations were all small *and negative*. MI score was another apparently

negligible variable (see also Durrant, 2014); whereas concreteness-imageability accounted for 7.7% of the variation and assonance 3.3%.

Three meta-analyses

Introduction

A statistical meta-analysis (i.e., quantitative research synthesis) is a principled approach to synthesizing the results of an original study plus one or more replications. A statistical meta-analysis can have at least two goals. One of these is to provide an authoritative conclusion about the direction, size, and substantive importance of a focal effect or set of effects (e.g., Borenstein, Hedges, Higgins, and Rothstein, 2009). Another goal is that of furnishing an interim estimate of a hypothesized effect's direction and size so that members of a research community can better judge whether it merits further investigation and so that prospective investigators may be better informed about the likely effect size when carrying out a priori estimation of statistical power for the purpose of estimating, say, how many experimental participants a replication would require (Anderson & Maxwell, 2016; Braver, Theommes, & Rosenthal, 2014; Cumming, 2012, 2014; Mullen, Muellerleihe, & Bryant, 2001; Rosenthal, 1990). As indicated by Braver et al. (2014), meta-analyses that are informed by this second goal are comparatively prospective, as opposed to summative. Importantly, they may include many fewer studies than meta-analyses of the first type, and they may repeatedly be extended by adding in new studies in order to gain an increasingly precise estimate of ES (e.g., Braver et al., 2014; Cumming, 2012, 2014; Mullen et al., 2016). Such 'continuously cumulating meta-analysis' (Braver et al., 2014) is likely to be especially beneficial in L2 research due to its heavy reliance on small samples of learners and of vocabulary items (Lindstromberg & Eyckmans, 2017). In fact, if a fixed-effects approach is adopted (see further below) it can be advantageous to meta-analyze as few as two studies due

to the potential of meta-analysis to improve the precision of estimates and simultaneously increase statistical power (Borenstein et al., 2009; Cumming, 2012): For an example of a two study meta-analysis in applied linguistics see Ellis and Sagarra (2011). Each of the new meta-analyses we report below include *eight* primary studies. The motive for conducting them was to add to the existing sparse evidence about the approximate size and direction of frequency effects on the development of productive knowledge of L2 MWEs in situations of deliberate learning.

The purpose and nature of the previously conducted primary studies

The three small-scale meta-analyses we report concern the following eight quasi-experimental studies, hereafter referred to as the 'primary studies'. In more or less chronological order they are: Boers, Lindstromberg, & Eyckmans (2012); Boers, Lindstromberg, & Eyckmans (2014b), studies 1 and 2; Boers, Lindstromberg, & Eyckmans (2014c), study 2; Lindstromberg & Boers (2008), studies 1 and 2; Lindstromberg & Eyckmans (2017); and finally, a pilot study for study 1 of Lindstromberg & Boers (2008). All these studies were carried out for a different purpose than that of estimating frequency effects. However, data collected in these studies have potential to cast light on such effects when they are reanalyzed. To elaborate, all eight studies were designed to investigate hypothesized positive mnemonic effects of interword, intraMWE patterns of phonological similarity, or sound repetition (SR)–for example, alliteration as in *full force*, rhyme as in *deep sleep*, and assonance as in *town house*. The treatment phase of each study included researcher-led direction of attention to phonological and orthographic form and, in some cases brief researcher-led awareness-raising about SR (see Table 1 for an overview). In all of these studies each targeted MWE consists of two words, mostly with the structure Adj.-N or N-N. As explained in more detail in the next section, in each study the set of targets included sound-repeating (SR) MWEs and nonSR control MWEs. The underlying research hypothesis

of these studies was that the forms of deliberately studied SR MWEs would be easier to freely recall than nonSR MWEs, on the assumption that free recall of the SR MWEs would be facilitated by phonological priming to a much greater degree. These eight studies are the only ones chosen for meta-analysis in the current study simply because they are the only ones we were able to find that addressed the retrievability of deliberately studied L2 MWEs and which did so with the focus on free recall of *forms*. We excluded studies where the focus was on cued recall of forms or on free recall of forms *plus* recently learned form-meaning mappings (a situation that may occur when learners are tested on recently studied MWEs that were previously unknown). It may be noteworthy that one relevant published study could not be included because the surviving records include only the per-learner posttest scores, not the essential per-MWE posttest scores.

Table 1.

Overview of the eight original studies.

Study[a] $N_{Learners}$ & $N._{MWEs}$	*L1 / Treatment*
(1) L & B (2008), #1. 25 & 26	L1: Dutch. / The learners (Ls) were already familiar with alliteration. The 26 MWEs were shown on jumbled cards. (There were 26 cards per set). (1) Dictation of the MWEs in pairs/threes: Ten Ls dictated; 15 wrote the MWEs down. (2) Individually, each L sorted a set of jumbled cards into a pile for alliterative MWEs (Allits) and a pile for nonalliterative MWEs (nonAllits).
(2) Pilot for the study above. 6 & 22	L1: German or Russian. / The procedure was the same as above.
(3) L & B (2008), #2. 31 & 24	L1: Dutch. / Similar to the above but with no dictation. The Ls each sorted 24 cards, each showing one MWE that rhymes, alliterates, assonates, or is nonSR. There were six MWEs of each type. The task was to sort the MWEs four ways according to sound pattern. Beforehand, there was general awareness-raising about pleasing patterns of sound repetition but not about specific patterns.
(4) B, L, & E (2012).	L1: Malaysian or Chinese. / The teacher called out the focal MWEs (matched Allits and nonAllits) plus nonAllit filler MWEs in jumbled order. The Ls

27 & 30	chorally recited each MWE and then wrote it down. There was no awareness raising about SR.
(5) B, L & E (2014a), # 1. 55 & 20	L1: Malaysian or Chinese. / The procedure was essentially the same as for Study 4 above except that now half the targeted MWEs were assonant rather than alliterative.
(6) B, L & E (2014a), #1 44 & 20	L1: Dutch. / The procedure was the same as for Study 6 above except that, additionally, there was prior awareness raising about assonance and, just following the dictation the learners were asked to identify the MWEs that show assonance.
(7) B, E, & L (2014), #2. 47 & 28	L1: Dutch. / The teacher called out each of the 28 focal MWEs twice, along with fillers. The 47 Ls were to repeat each MWE subvocally, locate it on a jumbled list, and assign it a 2, 1, or 0 according to whether they thought they heard, read, or used the MWE often, sometimes, or (almost) never. Then Ls dictated the MWEs to each other.
(8) L & E (2017). 81 & 28	L1: Dutch. / The teacher dictated the targeted MWEs in jumbled order. Following this, there was awareness-raising about assonance. Then learners were asked to examine the dictated MWEs and mark the ones that do *not* assonate.

Note. B = Boers; E = Eyckmans; L = Lindstromberg.

Learners, stimulus expressions, and procedures in the original studies

Learners. The 406 learners who participated in the eight primary studies were all adults, of whom nearly 80% were L1 Dutch university students around 20 years of age. All the learners were at least at B2 level according to the Common European Framework of Reference for Languages (CEFR), which corresponds to an IELTS score of 5 to 6.5. Note that 406 is about seven times the median sample size in L2 (quasi)experimental studies with a between groups or mixed design (Lindstromberg, 2016, pp. 746-747).

Stimulus expressions. In each of the studies the to-be-learned MWEs were selected nonrandomly from larger lists of candidate MWEs. The goal was always to screen out MWEs that are conspicuously idiomatic, rare, obscure in meaning or associated only with a particular variety of English. Also excluded were MWEs containing a word judged to be highly emotive (e.g., *sex, death*) or a word denoting an animal. These exclusions were

implemented in order to minimize the possibility of floor or ceiling effects among the eventual per-MWE test scores. (Regarding the superior memorability of emotive and animate words see, respectively, Schmidt, 2011, and VanArsdall, et al., 2013). An additional goal when selecting to-be-learned MWEs was to achieve an approximate overall balance between the SR and the control MWEs in terms of syllable count, types of syllable structure, constituent word frequency, concreteness-imageability rating, and (in some cases) mutual information score. Concreteness is an especially important predictor of the ease with which lexical items can be recalled (e.g., West & Holcomb, 2000). Concreteness correlates so strongly ($r \approx .80 - .90$) with another important variable, imageability of meaning (Paivio, 1986), that in some of the primary studies (and in our new meta-analytic study as well) we followed many other researchers (e.g., Dellantonio, Mulatti, Pastore, & Job, 2014; Reilly & Kean, 2007) in using just one of these variables to stand for both.

Finally, across the eight studies, 139 different MWEs were targeted (see the Appendix). Of these, 97 (70%) were used in only one study, 21 (15%) were used in two studies, 19 in three studies (13.5%), and two (1.5%) in four studies. To put the total of unique MWEs in perspective, in a survey of L2 (quasi)experimental research reports appearing in one journal over 20 years, Lindstromberg (2016, p. 747) found that 15 was the median number of linguistic items targeted in the studies in which dependent measures such as test scores were aggregated (i.e., subtotalled) by vocabulary items. Our overall sample of unique MWEs is more than nine times larger than this.

Procedures. In the treatment phase of each of the primary studies the participating learners encountered all the to-be-learned MWEs either in a random order or in a quasi-random order that avoided juxtaposition of any matched-MWEs such as *town house*$_{.\mathrm{SR}}$ and *town square*$_{.\mathrm{nonSR}}$. In most of the studies there was an interlude between the treatment and the immediate posttest during which learners were asked to engage in an unrelated activity

lasting a minute or more. The purpose of this was to disrupt attempts by learners to mentally rehearse the studied MWEs. As mentioned, Table 1 (above) provides an overview of the treatments.

A new meta-analytic study

Introduction

In our reanalysis of the data from the primary studies, the dependent variable remained the productive knowledge of form as measured by the per-MWE free recall scores. We focused on the immediate and near immediate posttests (hereafter, referred to as 'immediate posttests' or '(near) immediate posttests') rather than any delayed posttests because some of the studies either had no delayed posttest or else the delayed test was a test of cued recall or of recognition. The focal explanatory variable was frequency, as measured by (a) whole MWE frequencies, (b) per-MWE total constituent word frequencies, and (c) frequencies of the least frequent word in each MWE. To illustrate the latter measure, *speed* is the least frequent word in the expression *high speed.* Our rationale for including this last measure is as follows. On the basis of past experience we expected LF words to be more readily recalled after a session of deliberate learning than HF words. Thus, we expected the *MWEs* containing words of comparatively low frequency to be comparatively well recalled as well. The theoretical grounds for our expectation are as follows. In currently dominant connectionist models of language production (e.g., Dell & Gordon, 2003), people's mental representations of words that they have experienced in close association are likely to be neurologically linked, whether strongly or weakly. Given interword associations of this kind, recalling any given word in a cluster of associated familiar words has the potential to trigger recall of another word in the cluster through spreading activation (e.g., Loftus & Loftus, 1974). What is relevant here is that the constituent words of a MWE are ipso facto associated for anyone who knows the MWE. As might therefore be expected, it has been found that one

word in a familiar two-word MWE can serve as an effective retrieval cue for the other word, although nouns tend to be better cues than, say, adjectives while concrete words tend to be much better cues than abstract words, which may in fact have negligible cuing potential (Begg, 1972).[2] So, we speculated that if studied LF words are more likely to be recalled than studied HF words (for reasons outlined in the literature review further above), then the following scenario is somewhat more likely to play out in the case of LF words than in the case of HF words: First, one constituent word of a MWE is immediately recalled. Having been recalled, it can then act as a retrieval cue for the other word, if that word was not immediately recalled as well (cf., Sonbul & Schmitt's, 2013, discussion of 'collocation priming'). If this scenario actually happens, it could account for a negative correlation between frequency and MWE recall.

Two potential moderating explanatory variables were also of concern. One of these was sound repetition (SR), which was dummy coded 1 for 'present' or 0 for 'absent'. We thought it necessary to include this variable in our regression models owing to the fact that it had been the focal explanatory variable in the primary studies. The second potential moderating explanatory variable was concreteness. It was included especially because LF English words tend to be somewhat more imageable (and therefore more concrete) than otherwise similar English HF words (e.g., Stadthagen-Gonzales & Davis, 2006). We focused on concreteness rather than imageability because of the availability of a list of 40,000 recently collected concreteness ratings for word and MWE lemmas compiled by Brysbaert, Warriner, and Kuperman (2014; see http://crr.ugent.be/archives/1330). In this list we were able to find ratings for about 20% of the MWEs targeted in the eight primary studies. To obtain ratings for the remaining MWEs we used the mean of the ratings given for the two constituent words of each one. As it happens, BWK's list gives no ratings for a few of the constituent words. So in each of these cases we used the rating given for a related word or a

synonym. Generally, the alternative word differs from the target word only in part of speech. For example, one targeted MWE is *good guess*; there being no rating for *guess* (noun), we used the rating for *guess* (verb). In two cases though we had to use the rating of a synonym. So, for *kind* (adjective) we used the rating given for *kind-hearted* and for *use* (noun) we used the rating for *function* (noun).

Returning now to frequency, we used the values (obtained from COCA) that were given in the original reports. In cases where frequencies were not reported, we used COCA to get them anew for the present study. As pointed out by a reviewer, it is possible that frequencies might not be commensurable from one study to another because even though each individual study used frequencies collected from COCA at about the same time (usually within the same week), the time (e.g., the year) was different for different studies. This could matter because COCA grows by 20 million words a year. Due to COCA's policy of maintaining about the same genre balance from year to year (Davies, 2017), one might expect relative word frequencies to be quite stable from year to year. To check, we used the statistical freeware R (R Core Team, 2018) to randomly sample 150 items from COCA's 5000 most frequent word lemmas. We then used COCA's online search facility to find the frequencies of each of those items (more exactly, the frequencies of the corresponding base forms) in the sub-corpora that were added to COCA during six multi-year periods from 1990 through 2017. We chose the 5000 lemma list because it covers almost all of the words appearing in the MWEs targeted in the eight primary studies. The correlations shown in Table 2 indicate that relative frequencies of words in the 5000 lemma list have tended to be stable. We therefore conclude that it is safe to meta-analyze studies that used COCA frequencies from different years.

Table 2.

Correlations between COCA word frequencies in sub-corpora collected in adjacent spans of years.[a]

The periods compared →	*1990-94 and 1995-1999*	*1995-99 and 2000-2004*	*2000-2004 and 2005-2009*	*2005-2009 and 2010-2014*	*2010-2014 and 2015-2017*
Pearson's r on logged frequencies	.938 .855—.985[b]	.984 .971—.992[b]	.989 .982—.994[b]	.991 .986—.995[b]	.919 .834—.973[b]
Spearman's r	.92	.97	.98	.98	.89

Notes.

a. The first five periods span five years; the sixth spans only three years.

b. Bootstrapped 95% confidence interval (10000 iterations).

Research questions

The present study was intended to address the following questions with respect to each of our three measures of frequency (i.e., whole MWE frequency, MWE total constituent word frequency, frequency of the least frequent word in each MWE) and its observed effect on (near) immediate free recall of the forms of studied MWEs:

(1) Is the effect positive or negative?

(2) How big is it?

(3) Does it seem big enough to have substantive significance?

Owing to our previous experience in investigating the deliberate learning of L2 MWEs and owing to our readings in the literature on frequency effects on the item recall, we expected to observe a small to moderate negative effect. That said, we were mindful that disparate results have been reported in the literature. We were mindful too of the fact that only eight of the targeted MWEs include a word lemma that is outside the 5000 most frequent word lemmas in COCA. (The least frequent of these words is *sadly*.) Therefore, the constituent words figuring in the primary studies might not fully display the unusual

structural characteristics of LF words that tend to attract learners' attention most strongly, meaning that any moderating effects of novelty and distinctiveness could be attenuated.

Data analysis: Final steps before the meta-analysis

Following normal practice (Baayen, 2008) we converted our sets of frequencies to (natural) logarithms. (Table 3 shows how the three frequency measures correlate in our data.) Then, proceeding study by study, we used R to test a simple additive multiple linear regression model for each primary study. As already indicated, the outcome variable in this model is *Recall* while the explanatory variables are *Frequency* (the focus) along with *Concreteness* and *SR* (sound repetition) as covariates. In two of the eight primary studies, some of the constituent words were polysyllabic. For these studies, we also ran models with Number of Syllables as covariate; but we eliminated this variable during the model comparison stage of data analysis owing to its lack of statistically and substantively significant explanatory power. For our regression models, the measure of the size of the effect of frequency was its semi-partial correlation, r_{SP}.[3] Each value of r_{SP} for Frequency expresses the correlation between Recall and that part of Frequency that is independent of Concreteness and SR. (For further discussion of r_{SP} and for details of its use in meta-analysis see Aloe & Becker, 2012).

Table 3.

Pearson's correlations between the three sets of logged frequencies for the 139 unique multiword expressions (MWEs)

Frequency Measures ↓ →	Whole MWE	Total word frequency
Both words combined	.54	--
Least frequent word	.51	.66

As already stated our measure of form recall was the per MWE test scores, that is, the number of learners who had recalled each MWE. Formerly, the use of such 'by item' scores was standard practice; but a better, contemporary approach involves use of mixed-effects logistic multiple linear regression (Baayen, 2008). This superior option was not open to us because the surviving records for a few of the early studies were insufficiently detailed. Specifically, only aggregated scores survived, meaning that we knew how many learners had recalled each MWE and how many MWEs had been recalled by each learner but we could not tell *which* learners had recalled *which* MWEs. Our need to rely on scores aggregated by MWEs left us with a technical problem that mixed-effects regression automatically solves. To explain, key inputs for the meta-analyses were, for each study, the observed value of r_{SP} from each study and the sample size. One might at first think the sample size would be equivalent to the number of targeted MWEs (*Mdn* = 25). However, the eight primary studies involved quite different numbers of learners (*Range*: 6-81, *Mdn* = 37.5). To illustrate, two of the eight studies targeted 24 MWEs, namely, Lindstromberg and Boers (2008, Study 1) and Lindstromberg and Eyckmans (2017). However, the first of these studies involved 25 learners whereas the second involved 81. The estimate of r_{SP} stemming from the study with 81 learners should carry more weight in a meta-analysis for at least two reasons. First, all else being equal, that study would have the most statistical power and would accordingly yield the most precise and credible estimate of r_{SP} in the population. Second, in studies with our experimental design statistical power is largely a function of the *product* of $n._{MWEs}$ and $n._{learners}$ (Westfall, Kenny, & Judd, 2014). So what is wanted is a way to reduce the original two-fold sample size of each study to a single number that takes account of the relation between this product and statistical power. The problem here is analogous to the problem faced by someone wanting to find a good way to compare the sizes of fields that have disparate rectangular shapes. As we know, a common way to do this is to calculate and

compare the fields' areas. One can go further and re-express each rectangular field as a square field of equivalent area. Then, a side of the square is equal in length to the square root of the area. From mathematics classes we may remember that the number representing this side-length is in fact the *geometric mean* (GM) of the height and width of each original rectangle since $GM_{x\&y}$.= sq. root(x times y). These GMs can be used to rank the fields by size in a way that takes the differing *products* of the original heights and widths into account. Of course, the fields could also be ranked by area; but areas are measured not in lengths but in *squared* lengths (e.g., square meters), which may not be convenient for some purposes. So, following the gist of the procedure just outlined, we calculated the GM of each study's $n._{\text{MWEs}}$ and $n._{\text{learners}}$ and fed the GM into the software as the study sample size. To illustrate, the GM of $n._{\text{MWEs}} = 24$ and $n._{\text{learners}} = 25$ is 24.4949 and the GM of $n._{\text{MWEs}} = 24$ and $n._{\text{learners}} = 81$ is 44.09082, whereby the latter study is upweighted compared to the former one—but not as radically as would happen if the arithmetic mean or median were used on the false assumption that there is an *additive* relation between statistical power and $n._{\text{MWEs}}$ and $n._{\text{learners}}$. For example, the mean and median of 24 and 25 are both 24.5 whereas for 24 and 81 they are both *52.5*.[4]

The meta-analytic approach

To run the meta-analyses, one for each measure of frequency, we used the R package *metafor* (Viechtbauer, 2010), which implements approaches to meta-analysis described by Borenstein et al. (2009). Following the recommendation of Borenstein et al. we used random-effects (RE) meta-analysis which, unlike fixed-effects (FE) meta-analysis, allows for the possibility that the magnitude of the true effect of frequency might be different in studies in which learners experienced different conditions of attention direction. A RE meta-analysis tends to yield an overall estimate of ES that has a wider CI than the CI of a corresponding FE meta-analysis, for which reason RE meta-analysis may be regarded as comparatively

conservative. A caveat is that for RE meta-analysis it seems advisable to have at least six primary studies (Borenstein et al.; see also Guolo & Varin, 2017). Because our meta-analysis covered eight primary studies, which is still a relatively small number, we used an RE approach that includes a correction which controls the Type I error rate when the number of primary studies is, say, less than ten.[5]

Results of the meta-analyses

Notwithstanding our choice of RE meta-analysis, we expected the observed effects of frequency to be similar from study to study given that the eight studies used essentially the same design to address the same or similar research questions, even though the attention direction tasks were different in some cases. The statistics given in Table 4 indicate that our expectation was borne out, especially with respect to the effect of the frequency of the least frequent word. For example, the values of I^2 shown in the middle of the table are interpretable as measures of the between-study heterogeneity of observed effects. For values of I^2 the following benchmarks have been suggested: 25% = *low*, 50% = *moderate*, 75% = *high* (Borenstein et al., 2009, pp. 117-122). Thus, our observed levels of heterogeneity ranged from extremely low to moderate. It seems noteworthy here that the effect of whole MWE frequency was the most heterogeneous, that is, the least consistent. This is perhaps unsurprising given the potential for the processing of MWEs to be more complex than the processing of single words. Lastly, the Q statistic (Table 4, right column) is used for testing the null hypothesis that the primary studies share a common effect size. The three nonsignificant *p* values indicate that this hypothesis is neither rejected nor accepted.

Table 4.

Diagnostics of the consistency and heterogeneity of observed effects in the eight primary studies

Random-Effects Model (k = 10; tau² estimator: REML)		
Index of Frequency	I^2	Q, $df = 7$
Least Frequent Word	0%	4.31, $p = .74$
Total Words	36.1%	10.61, $p = .16$
Whole MWE	44.8%	12.26, $p = .09$

We now turn from diagnostic statistics to the key results of the three meta-analyses, beginning with the meta-analysis having to do with the effect of the least frequent word. Let us focus first of all on Figure 1, which includes graphical representations of the eight 95% confidence intervals (CIs) for the eight values of r_{SP}. A notable feature of these graphics is that the area of the box in the center of each CI is proportional to the weight given to that value of r_{SP} in the meta-analysis. The smallest box and, accordingly, the widest CI pertains to the pilot study involving just six learners. Another notable feature is the laterally stretched black diamond in the bottom part of the figure: This represents the CI for the overall averaged, or pooled, estimate of *rho* which, given our RE approach, is the *mean* of the various potentially different true effects (i.e., the various true semi-partial correlations) that are associated with the various ways of directing learners' attention to phonological forms. If we square the estimate of *rho*—which is -.14 or, more exactly, -0.138845—we get $r_{SP}^2 = .019$ → 1.9% as an estimate of the percentage of variation in free recall that is accounted for by the frequency of an MWE's least frequent word following some kind of direction of attention to phonological forms. While this percentage is indicative of an effect which will be of minor

practical significance in many situations (cf., Ellis, 2010, p. 41), in a context such as vocabulary learning where effects are ongoing and likely to accumulate an effect that is numerically much smaller than this can have appreciable *practical* significance (Abelson, 1985). We should note though that the range of effects indicated by the CI is quite wide: The value of r_{SP}^2 corresponding to the 'left' limit of the CI equates to 5.7% of variation explained (indicative of a fairly solid practical significance) while the value of r_{SP}^2 corresponding to the other limit of the CI equates to just a bit more than 0.1%.

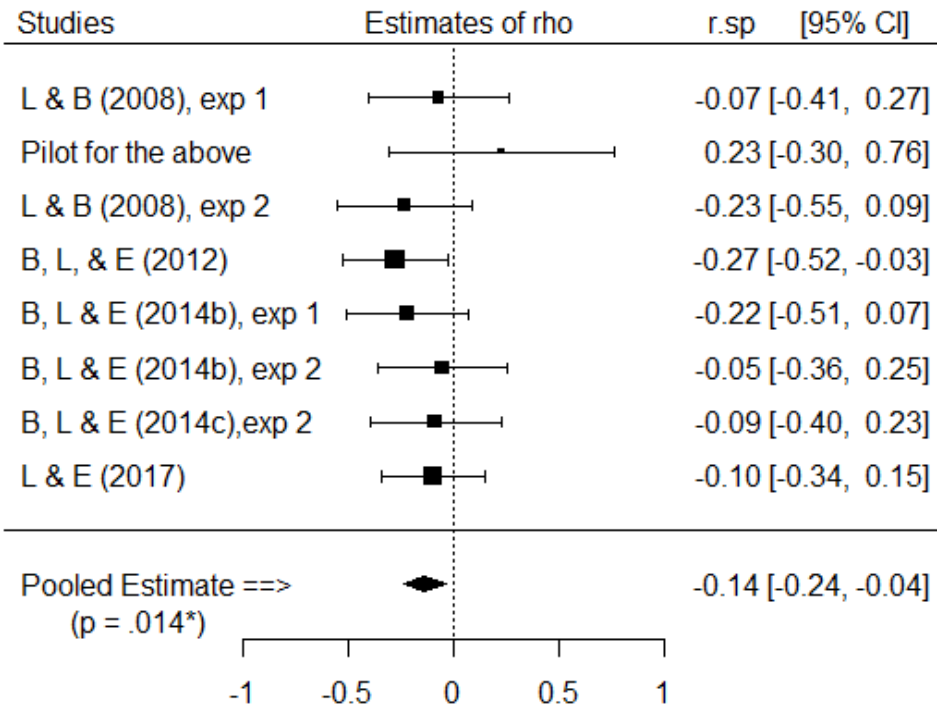


Figure 1.

A forest plot giving an overview of the results of the meta-analysis relating to the (logged) frequency of the least frequent constituent word in a MWE. The measure of effect size (ES) is the semi-partial correlation, r_{SP}, and *rho* is the mean population semi-partial correlation that is being estimated. Note also that *B* = Boers, *E* = Eyckmans, *L* = Lindstromberg, and *exp* = experiment.

Statistics for the other two meta-analyses are given in Table 5. As can be seen, the pooled estimate of ES for total constituent word frequency is almost the same as the estimate for frequency of the least frequent word (see Figure 1). In contrast, the small positive estimate for whole MWE frequency corresponds essentially to a null effect (r_{SP}^2 = .001). It might be deemed prudent to control of our overall Type I error rate by means of, say, a Bonferroni-type correction such as Hochberg's procedure. If this were done, only the *p* value shown in Figure 1 would remain significant. However, such procedures are known to be overly conservative when significance tests are positively correlated (e.g., Conneely & Boehnke, 2007), which our three tests undoubtedly are (as can be inferred from Table 3). Finally, it may be of interest that in the regression models for the primary studies, Concreteness virtually always had much more explanatory power than either SR or Frequency, whose observed ESs were almost always in different directions from each other and usually $p > .05$.

Table 5.

Statistics for the meta-analyses of the effects of total constituent word frequency (TCW) and whole MWE frequency. The semi-partial correlation is the measure of effect size (ES). The outcome variable is (near) immediate free recall.

Measure of frequency	*Pooled estimate of ES*	*Standard error*	*95% CI of the Estimate*	*z*	*P*
TCW	-.16	0.07	-.31 ↔ -.004	-2.42	.046*
Whole MWE	.04	0.07	-.13 ↔ .20	0.53	.617

Summary and discussion

The study reported above casts light on an important but little investigated matter—the effect of frequency on L2 learners' ability to recall MWE *forms* in a situation of

deliberate learning. Our three research questions concern the *direction* of any such frequency effect, its *size*, and its *practical importance*. Our measures of frequency were the frequency of the least frequent word (LFW), total constituent word (TCW) frequency, and whole MWE frequency. Our conclusions regarding our research questions differ somewhat depending on the frequency measure. Going by TCW frequency and, especially, LFW frequency, we conclude that the effect of frequency on MWE recall is negative and rather small but that it may nevertheless be large enough have practical significance given that the effect could operate in many encounters with MWEs throughout years of learning. Our results for whole MWE frequency are quite vague: The point estimate is trivially positive, but the confidence interval (or CI)—about 60% of which lies above zero—includes both negative and, especially, positive effect sizes that could conceivably have practical significance. As to the three measures of frequency, it appears that the most sensitive one for the purpose at hand is LFW frequency and that whole MWE frequency is the least sensitive. This conclusion is indicated partly by the fact that the CI for the estimate of the pooled effect of whole MWE frequency straddles zero but most especially by the fact that this CI is 65% wider and correspondingly less informative than the CI for estimated pooled effect of LFW frequency. (Prior evidence suggests that LFW frequency should be the better measure anyway: Recall our earlier discussion of findings that comparatively infrequent words attract a disproportionate amount of attention.) Turning now to TCW frequency, it seems that as a measure it may at best be redundant for the purpose it served in our study. For one thing TCW frequency is strongly correlated with LFW frequency (see Table 3). For another, the CI for its overall estimate of effect is also comparatively wide, being 54% wider than the corresponding CI for the effect of LFW frequency.

Limitations

We must stress that our conclusions must be regarded as tentative owing to the limitations of our study. First, all three of our meta-analyses are small in scale. Second, the primary studies should also be regarded as small in scale even though the average sample sizes of learners and linguistic items were above average for L2 research. Third, none of the primary studies featured random selection of the to-be-learned MWEs from a large pool of candidate MWEs representative of MWEs that the participating learners were likely to know. This absence of randomization means that there is no way of assessing the extent to which our results may have been biased by variables (apart from concreteness and SR) with known or likely potential to influence retrieval of lexical items from episodic memory—for instance, bigram frequency, phonological neighbourhood density, and cognateness (De Groot & Keijzer, 2000), to mention just three of many uncontrolled variables. This limitation is of course inherent to quasi-experimental studies. Fourth, 20% of the 139 unique stimulus MWEs were used in more than one study. Although the resulting repeated measurements may have led to some increased measurement reliability, our estimates of effect size are bound to be less generalizable than they would be if each study had used entirely different MWEs. Fifth, our traditional, by-items approach to the statistical analysis of test scores restricts the possibility of generalizing to MWEs other than the ones actually targeted in the primary studies (Baayen, 2008). Regarding what might appear to be a sixth limitation, we should reiterate that our study is based on data from primary studies involving MWEs consisting of words that participating learners were highly likely to have already learned, receptively at least. The rationale for targeting such expressions in a study of form retrievabilityl has to do with evidence that learners generally lack sufficient processing capacity to focus on forms and on new meanings or on new form-meaning mappings at the same time (Barcroft, 2015). Thus, targeting expressions whose meanings are known may permit retrievability of their forms to be investigated against a less noisy background. This was a key consideration also in

the primary studies where the focal explanatory variable was the presence/absence of a pattern of interword sound repetition, which is a type of *form* variable. As to the *current* study, a second rationale for targeting MWEs made up of words that learners are likely to know at least receptively is that (for reasons already outlined) any given learner may not yet have developed productive knowledge of many potentially useful MWEs composed of familiar, relatively high frequency (HF) words: We have referred to evidence that HF words, and therefore probably also the MWEs in which they occur, tend to be given comparatively little attention when encountered in ordinary input. As noted too, MWEs made up of HF words may be hard to learn because these words are particularly likely to be confusable (Martinez & Murphy, 2011). For one thing, many HF words are polysemic with common delexicalized uses, and their meanings tend to be of comparatively low imageability (e.g., Tellings, et al., 2013). Both of these characteristics are likely to be associated with common learner errors of word choice—as in *do a mess** and *make an experience**, where the relevant verbs (*do*, *make*, and also *have*) are all polysemic and relatively unimageable as well. While L1 influence is often the likeliest origin of such errors, polysemy and low imageability may hinder eventual adoption of conventional forms even when these forms occur quite frequently in input. For further discussion of these and related issues see especially Boers et al. (2014) and Boers, Dang, & Strong (2017).

Conclusion

As mentioned, we carried out the three meta-analyses in order to provide *interim* estimates of effect size. The ideal way to obtain more precise estimates of the relevant frequency effect(s), would be to carry out one or more relatively large-scale studies that are free from the limitations of the eight primary studies upon we based our meta-analyses. In particular, targeted MWEs should be selected entirely at random from a very large pool of candidate MWEs. If it were established that frequency has a weak positive effect or else a

negative effect on ability to recall the form of a MWE, there would be at least two implications. The first concerns HF MWEs made up of HF words: MWEs of this kind which are not yet known productively would be prime targets for extra pedagogical attention not only on account of their probable usefulness but also because of the likelihood that their HF status could not be relied on to facilitate form acquisition in the absence of extra pedagogical attention. Second, if there is indeed a low frequency (LF) advantage in form recall, there could also be a practical advantage in increasing the targeting of MWEs that include one or more LF words on the grounds that productive knowledge of these MWEs may be gained relatively quickly, provided of course that the *meanings* of the MWEs and their constituent words are not difficult to learn beforehand. If this latter course of action were adopted, it would be reasonable to prioritize those LF-word MWEs which have one or more additional characteristics known to facilitate acquisition, for example, animacy (VanArsdall et al., 2013), cognateness (De Groot & Keijzer, 2000), a high degree of concreteness (De Groot & Keijzer, 2000), enactability (Asher, 1969; Cohen, 1989), high imageability (Paivio & Desrochers, 1979), and the presence of interword sound repetition (e.g., Boers & Lindstromberg, 2009). With specific respect to forms, regardless of any frequency effect it makes sense for teachers to alert learners to the presence of any facilitative formal characteristics that an appreciable proportion of learners will probably overlook—for example, cognateness and the presence sound repetition.

Acknowledgements

We are very grateful to the reviewers for helping us improve this article and to Frank Boers for valuable help with an intermediate draft.

Notes

1. A ‘base word’ is the uninflected form of a word (e.g., *hit* as opposed to *hits* and *hitting*). A ‘lemma’ is base word *and* its inflected forms. Thus, the frequency of a lemma cannot be less than the frequency of its corresponding base word. As with a single word, the frequency of a MWE can be measured in terms of its lemma or in terms of any of its of fixed forms, such as its canonical base form.
2. See Durrant and Doherty (2010) for a relevant discussion of types of association between constituent words in L1 and L2 MWEs.
3. Consider a regression model with three explanatory variables (EVs) A, B, and C. We can calculate r_{SP} for variable A, for example, as follows: We test the full model that includes all three EVs A, B, and C and also a model that includes A and B but not C. We take the value of R^2 for the second model and subtract it from the value of $R^2$for the full model. The result is r_{SP}^2 for variable A. This expresses the amount of variation in the dependent variable that is attributable to A over and above the variation attributable to B and C. To get r_{SP} we take the square root of r_{SP}^2.
4. The geometric mean (GM) is used for various purposes in a wide range of fields including medical research. For our purposes the GM has two advantages. First, it can be used to find an average of things that are different, such as weight and height. Second, the GM of two unequal positive numbers is always less than their mean or median, which can make the GM especially suitable when a lower or upper value is extreme. For nontechnical overviews see McChesney (2016) and http://www.statisticshowto.com/geometric-mean-2/.
5. This is the Knapp-Hartung correction. For discussion and positive evaluation, see Guolo & Varin (2017).

References

Abelson, R. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, 97(1), 129–133. http://www2.psych.ubc.ca/~schaller/528Readings/Abelson1985.pdf

Aloe, A., & Becker, B. (2012). An effect size for regression predictors in meta-analysis. *Journal of Educational and Behavioral Statistics, 37*, 278–297. https://doi.org/pdf/10.3102/1076998610396901

Anderson, M. (2009). Retrieval. In A. Baddeley, M. Eysenck, & M. Anderson (Eds.), *Memory*, pp. 161–189. Psychology Press: Hove, UK.

Anderson, S. & Maxwell, S. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods, 21*(1), 1–12. https://doi.org/10.1037/met0000051

Asher, J. (1969). The total physical response approach to second language learning. *Modern Language Journal*, 53(1), 3–17. https://doi.org/10.1111/j.1540-4781.1969.tb04552.x

Baayen, H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Baddeley, A., Eysenck, M., & Anderson, M. (2009). *Memory*. Hove, UK: Psychology Press.

Barcroft, J. (2015). *Lexical input processing and vocabulary learning*. Amsterdam: John Benjamins.

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., Herron, D.,…& Tzeng, O. (2003). Timed picture naming in seven languages. *Psychonomic*

Bulletin and Review, *10*(2), 344–380. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3392189/

Begg, I. (1972). Recall of meaningful phrases. *Journal of Verbal Learning and Verbal Behavior*, *11*(4), 431–439. https://doi.org/10.1016/S0022-5371(72)80024-0

Boers, F., Dang, C.T., & Strong, B. (2017). Comparing the effectiveness of phrase-focused exercises. A partial replication of Boers, Demecheleer, Coxhead, and Webb (2014). *Language Teaching Research*, *21*(3), 362–280. https://doi.org/10.1177/1362168816651464

Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. (2014). Gauging the effects of exercises on verb-noun collocations. *Language Teaching Research*, *18*(1), 54–74. https://doi.org/10.1177/1362168813505389

Boers, F., Eyckmans, J., & Lindstromberg, S. (2014). The effect of a discrimination task on L2 learners' recall of collocations and compounds. *International Journal of Applied Linguistics*, *24*(3), 357–369. https://doi.org/10.1111/ijal.12033

Boers, F., & Lindstromberg, S. (2009). *Optimizing a Lexical Approach to Instructed Second Language Acquisition*. Basingstoke, UK: Palgrave-Macmillan.

Boers, F., & Lindstromberg, S. (2012). Experimental and intervention studies of formulaic sequences in a second language. *Annual Rev. of Applied Linguistics*, *32*, 83–110. https://doi.org/10.1017/S0267190512000050

Boers F, Lindstromberg, S., & Eyckmans, J. (2012) Are alliterative word combinations comparatively easy to remember for adult learners? *RELC Journal*, 43(1): 127−135. https://doi.org/10.1177/0033688212439997

Boers, F., Lindstromberg, S. & Eyckmans, J. (2014a). Some explanations for the slow acquisition of L2 collocations. *Vigo International Journal of Applied Linguistics, 11*, 41–62. http://vialjournal.webs.uvigo.es/pdf/Vial-2014-Article2.pdf

Boers, F., Lindstromberg, S., & Eyckmans, J. (2014b). When does assonance make lexical phrases memorable? *European Journal of Applied Linguistics and TEFL, 3*(1), 93–107. https://core.ac.uk/download/pdf/55762736.pdf

Boers, F., Lindstromberg, S., & Eyckmans, J. (2014c). Is alliteration mnemonic without awareness-raising? *Language Awareness, 23*(4), 291–303. https://doi.org/10.1080/09658416.2013.774008

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. Hoboken, NJ: Wiley.

Braver, S., Theommes, F., & Rosenthal, R. (2014). Continuously accumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9*, 333–342. https://doi.org/10.1177/1745691614529796

Brown, R. (1976). *A first language: The early stages*. Harmondsworth, MA: Penguin.

Brysbaert, M., Warriner, A., & Kuperman, V. (2014). Concreteness ratings for 40,000 generally known English word lemmas. *Behavior Research Methods, 46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Cohen, R. (1989). Memory for action events: The power of enactment. *Educational Psychology Review, 1*(1), 57–80. https://doi.org/10.1007/BF01326550

Conneely, K., & Boehnke, M. (2007). So Many Correlated Tests, So Little Time! Rapid Adjustment of P Values for Multiple Correlated Tests. *American Journal of Human Genetics, 81*(6), 1158–1168. https://doi.org/10.1086/522036

Cop, U., Keuleers, E., Drieghe, D., & Duyck, W. (2015). Frequency effects in monolingual and biliingual natural reading. *Psychonomic Bulletin and Review, 22*(5), 1216–1234. https://doi.org/10.3758/s13423-015-0819-2

Criss, A., Aue, W., & Smith, L. (2011). The effects of word frequency and context variability in cued recall. *Journal of Memory and Language, 64*(2), 119–132. https://doi.org/10.1016/j.jml.2010.10.001

Crossley, S., Salsbury, T.,Titak, A., & McNamara, D. (2014). Frequency effects in second language acquisition: Word types, word tokens, and word production. *International Journal of Corpus Linguistics, 19*(3), 301-332. https://doi.org/10.1075/ijcl.19.3.01cro

Cumming, G. (2012). *Understanding the* n*ew statistics: Effect sizes, confidence intervals, and meta-analysis*. Hove: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7–29. https://doi.org/10.1177/0956797613504966

Davies, M. (2008–2018). *The Corpus of Contemporary American English (COCA).* http://www.americancorpus.org

De Groot, A., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of words concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning, 50*(1), 1–56. https://doi.org/10.1111/0023-8333.00110

Dell, G. & J. Gordon (2003). Neighbors in the lexicon: Friends or foes? In N. Schiller & A. Meyer (eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities*. New York: Mouton, 9–37. https://www.researchgate.net/publication/303260474_Neighbors_in_the_lexicon_Friends_or_foes

Dellantonio, S., Mulatti, C., Pastore, L., & Job, R. (2014). Measuring inconsistencies can lead you forward: Imageability and the x-ception theory. *Frontiers in Psychology*, *5*, 708. doi.org/10.3389/fpsyg.2014.00708

DeLosh, E., & McDaniel, M. (1996). The role of order information in free recall: Application to the word-frequency effect. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*(5), 1136–1146. https://doi.org/10.1037/0278-7393.22.5.1136

Del Re, A. (2014). compute.es, version 0.2-4. (R package). https://cran.r-project.org/web/packages/compute.es/compute.es.pdf

Dewhurst, S., Brandt, K., & Sharp, M. (2004). Intention to learn influences the word frequency effect in recall but not in recognition memory. *Memory and Cognition*, *32*(8), 1316–1325. https://doi.org/10.3758/BF03206322

Diana, R., & Reder, L. (2006). The low-frequency encoding disadvantage: Word frequency affects processing demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 805–815. https://www.ncbi.nlm.nih.gov/pubmed/16822148

Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, *6*(2),125–155. https://doi.org/10.1515/cllt.2010.006

Durrant, P. (2014). Corpus frequency and second language learners' knowledge of collocations: A meta-analysis. *International Journal of Corpus Linguistics*, *19*(4), 443–477. https://doi.org/10.1075/ijcl.19.4.01dur

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

Ellis, N., & Sagarra, N. (2011). Learned attention in adult language acquisition: A replication and generalization study and meta-analysis. *Studies in Second Language Acquisition 33*, 589–624. https://doi.org/10.1017/S0272263111000325

Ellis, P. (2010). *Essential guide to effect sizes.* Cambridge: Cambridge University Press.

Eyckmans, J., Boers, F., & Stengers, H. (2007). Identifying chunks: Who can see the wood for the trees? *Language Forum, 33*, 85–100.

Eyckmans, J., & Lindstromberg, S. (2016). The power of sound in L2 idiom learning. *Language Teaching Research*, *21*(3), 341–365. https://doi.org/10.1177/1362168816655831

Eysenck, M., & Eysenck, C. (1980). Effects of processing depth, distinctiveness, and word frequency on retention. *British Journal of Psychology*, *71*(2), 263–274. https://doi.org/10.1111/j.2044-8295.1980.tb01743.x

Gardner, M., Rothkopf, E., Lapan, R., & Lafferty, T. (1987). The word frequency effect in lexical decision: Finding a frequency-based component. *Memory and Cognition 15*(1), 24–28. https://doi.org/10.3758/BF03197709

Glanzer, M., & Ehrenreich, S. (1979). Structure and search of the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 18*(4), 381–398. https://doi.org/10.1016/S0022-5371(79)90210-X

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

Goldstein, R., & Vitevitch, M. (2014). The influence of clustering coefficient on word-learning: How groups of similar sounding words facilitate acquisition. *Frontiers in Psychology, 18*. https://doi.org/10.3389/fpsyg.2014.01307

Gordon, B., & Caramazza, A. (1982). Lexical decision for open- and closed-class words: Failure to replicate differential frequency sensitivity. *Brain and Language, 15*(1), 143–160. http://www.wjh.harvard.edu/~caram/PDFs/1982_Gordon_Caramazza.pdf

Guolo, A., & Varin, C. (2017). Random-effects meta-analysis: The number of studies matters. *Statistical Methods in Medical Research, 26*(3), 1500–1518. https://doi.org/10.1177/0962280215583568

Hall, J. (1954). Learning as a function of word-frequency. *American Journal of Psychology, 67*(1), 138–140. https://doi.org/10.2307/1418080

Hamrick, P., & Rebuschat, P. (2014). Frequency effects, learning conditions, and the development of implicit and explicit lexical knowledge. In J. Connor-Linton, & L. Amoroso (Eds.), *Measured language: quantitative approaches to acquisition, assessment, processing and variation*, pp. 125–139. Washington, DC: Georgetown University Press.

Han, M., Storkel, H., Lee, J., & Yoshinaga-Itano, C. (2015). The influence of word characteristics on the vocabulary of children with cochlear implants. *Journal of Deaf Studies and Deaf Education, 20*(3), 242–251. https://doi.org/10.1093/deafed/env006

Henriksen, B. (2012). Research on L2 learners' collocational competence and development. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *EUROSLA Yearbook 5: L2 vocabulary acquisition: Knowledge and use*, (pp. 29–56). Amsterdam: John Benjamins. http://www.eurosla.org/monographs/EM02/Henriksen.pdf

Hulstijn, J. (2001). Deliberate and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (ed.) *Cognition and Second Language Instruction*, pp. 258–286. Cambridge: Cambridge University Press.

Izura, C., & Ellis, A. (2002). Age of acquisition effects in word recognition and production in first and second languages. *Psicológica, 23,* 245–281. https://www.uv.es/revispsi/articulos2.02/4.IZURA%26ELLIS.pdf

Jenkins, J., Stein, M., & Wysocki, K. (1984). Learning vocabulary through reading. *American Educational Research Journal, 21*(4), 767–787. https://doi.org/10.3102/00028312021004767

Judd, C., Westfall, J., & Kenny, D. (2016). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68,* 17.1–17.25. https://doi.org/10.1146/annurev-psych-122414-033702

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavioral Research Methods, 44*(4),978–990. https://doi.org/10.3758/s13428-012-0210-4.

Laufer, B. 2005. Focus on Form in second language vocabulary acquisition. *EUROSLA Yearbook 5*(1), 223–250. https://doi.org/10.1075/eurosla.5.11lau

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Lindstromberg, S. (2016). Inferential statistics in Language Teaching Research: A review and ways forward. *Language Teaching Research, 20*(6), 741–768. https://doi.org/10.1177/1362168816649979

Lindstromberg, S., & Eyckmans, J. How big is the positive effect of assonance on the near-term recall of L2 collocations? *ITL International Journal of Applied* Linguistics, 165(1), 19–45. https://doi.org/10.1075/itl.165.1.02lin

Lindstromberg, S., & Boers, F. (2008). The mnemonic effect of noticing alliteration in lexical chunks. *Applied Linguistics* 29(2), 200–222. https://doi.org/10.1093/applin/amn007

Lindstromberg, S., & Eyckmans, J. (2017). The particular need for replication in the quantitative study of SLA: A case study of the mnemonic effect of assonance in collocations. *Journal of the European Second Language Association*, *1*(1), 126–136. https://doi.org/10.22599/jesla.26

Loftus, G., & Loftus, E. (1974). The influence of one memory retrieval on a subsequent memory retrieval. *Memory and Cognition*, *2*(3), 467–471. https://doi.org/10.3758/BF03196906

Lohnas, L., & Kahana, M. (2013). Parametric effects of word frequency in memory for mixed frequency lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(6), 1943–1946. https://doi.org/10.1037/a0033669

Lotto, L., & de Groot, A. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning, 48*(1), 31–69. https://doi.org/10.1111/1467-9922.00032

Martinez, R., & Murphy, V. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly, 45*(2), 267–290. https://doi.org/10.5054/tq.2011.247708

McChesney, J. (2016). You should summarize data with the geometric mean. https://medium.com/@JLMC/understanding-three-simple-statistics-for-data-visualizations-2619dbb3677a

McLeod, C., & Kampe, K. (1996). Word frequency effects in recall, recognition, and word fragment completion tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(1), 132–142. https://www.ncbi.nlm.nih.gov/pubmed/8648282

Meschyan, G., & Hernandez, A. (2002). Age of acquisition and word frequency: Determinants of object-naming speed and accuracy. *Memory & Cognition, 30*(2), 262–269. https://doi.org/10.3758/BF03195287

Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach.* Oxford: Clarendon Press.

Mullen, B., Muellerleihe, P., & Bryant, B. (2001). Cumulative meta-analysis: A consideration of indicators of sufficiency and stability. *Personality and Social Psychology Bulletin, 27*(11), 1450–1462. https://doi.org/10.1177/01461672012711006

Nation, P. (2013). *Learning Vocabulary in Another Language*, 3rd edn. Cambridge: Cambridge University Press.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology, 17*(4), 273–281. https://doi.org/10.1080/17470216508416445

Oxford University Press. (2017). How to use the *OED* key to frequency. https://public.oed.com/how-to-use-the-oed/key-to-frequency/

Paivio, A. (1986). *Mental Representations: A dual-coding Approach.* Oxford & New York: Oxford University Press.

Paivio, A., & Desrochers, A. (1979). Effects of an imagery mnemonic on second language recall and comprehension. *Canadian Journal of Psychology, 33*(1), 17–28. https://doi.org/10.1037/h0081699

Perfetti, C., & Hart, L. (2002). The lexical quality hypothesis. In L. Vehoeven. C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy,* (pp. 189–213). Amsterdam: John Benjamins. http://www.pitt.edu/~perfetti/PDF/Lexical%20quality%20hypothesis.pdf

Peters, E. (2014). The effects of repetition and time of post-test administration on EFL learners' form recall of single words and collocations. *Language Learning Research, 18*(1), 75–94. https://doi.org/10.1177/1362168813505384

Peters, E. (2016). The learning burden of collocations: The role of interlexical and intralexical factors. *Language Teaching Research, 20*(1), 113–138. https://doi.org/10.1177/1362168814568131

Pisoni, D., Nusbaum, H., Luce, P., & Slowiaczek, L. (1985). *Speech Communication, 4*(1–3), 75–95. https://doi.org/10.1016/0167-6393(85)90037-8.

R Core Team. (2017). R: A language and environment for statistical computing. (Version 3.3.1). Vienna: R Foundation for Statistical Computing. http://www.R-project.org/.

Reilly, J., & Kean, J. (2007). Formal distinctiveness of high- and low-imageability nouns: Analyses and theoretical implications. *Cognitive Science, 31*(1), 157–168. https://doi.org/10.1080/03640210709336988

Schmidt, S. (2011). Memory for emotional words in sentences: The importance of emotional contrast. *Cognition and Emotion, 26*(6), 1015–1035. https://doi.org/10.1080/02699931.2011.631986

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research, 12,* 329–363. https://doi.org/10.1177/1362168808089921

Siyanova-Chanturia, A., & Janssen, N. (2018). Production of familiar phrases: Frequency effects in native speakers and second language learners. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0000562

Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: Acquisition of collocations under different input conditions. *Language Learning, 63*(1), 121–159. https://doi.org/10.1111/j.1467-9922.2012.00730.x

Stadthagen-Gonzales, H., & Davis. C. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavioral Research Methods, 38*(4), 598–605. https://www.ncbi.nlm.nih.gov/pubmed/17393830

Storkel, H., & Morrisette, M. (2002). The lexicon and phonology: Interactions in language acquisition. *Language, Speech, and Hearing Services in Schools, 33*(1), 24–37. http://lshss.pubs.asha.org/article.aspx?articleid=1780270

Tellings, A., Coppens, K., Gelissen, J., & Schreuder, R. (2013). Clusters of word properties as predictors of elementary school children's performance on two word tasks. *Applied Psycholinguistics, 34*(3), 461–481. https://doi.org/10.1017/s014271641100083x

Tomasello, M. (2003). *Constructing a language*. Boston, MA: Harvard University Press.

VanArsdall, J., Nairne, J., Pandeira, J., & Blunt, J. (2013). Adaptive memory: Animacy processing produces mnemonic advantages. *Experimental Psychology, 60*(3), 172–178. https://doi.org/10.1027/1618-3169/a000186

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. http://www.jstatsoft.org/v36/i03/.

Waring, R. & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language, 15*(2), 130–163. http://nflrc.hawaii.edu/rfl/October2003/waring/waring.html

Watkins, M., LeCompte, D., & Kim, K. (2000). Role of study strategy in recall of mixed lists of common and rare words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(1), 239–245. https://doi.org/10.1037/0278-7393.26.1.239

Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning, 63*(1), 91–120. https://doi.org/10.1111/j.1467-9922.2012.00729.x

Westfall, J., Kenny, D., & Judd, C. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*, 220–245. https://doi.org/10.1037/xge0000014

Whaley, C. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior, 17*(2), 143–154. https://doi.org/10.1016/S0022-5371(78)90110-X

Appendix

THE 139 MULTI-WORD EXPRESSIONS TARGETED IN THE EIGHT PRIMARY STUDIES[a]

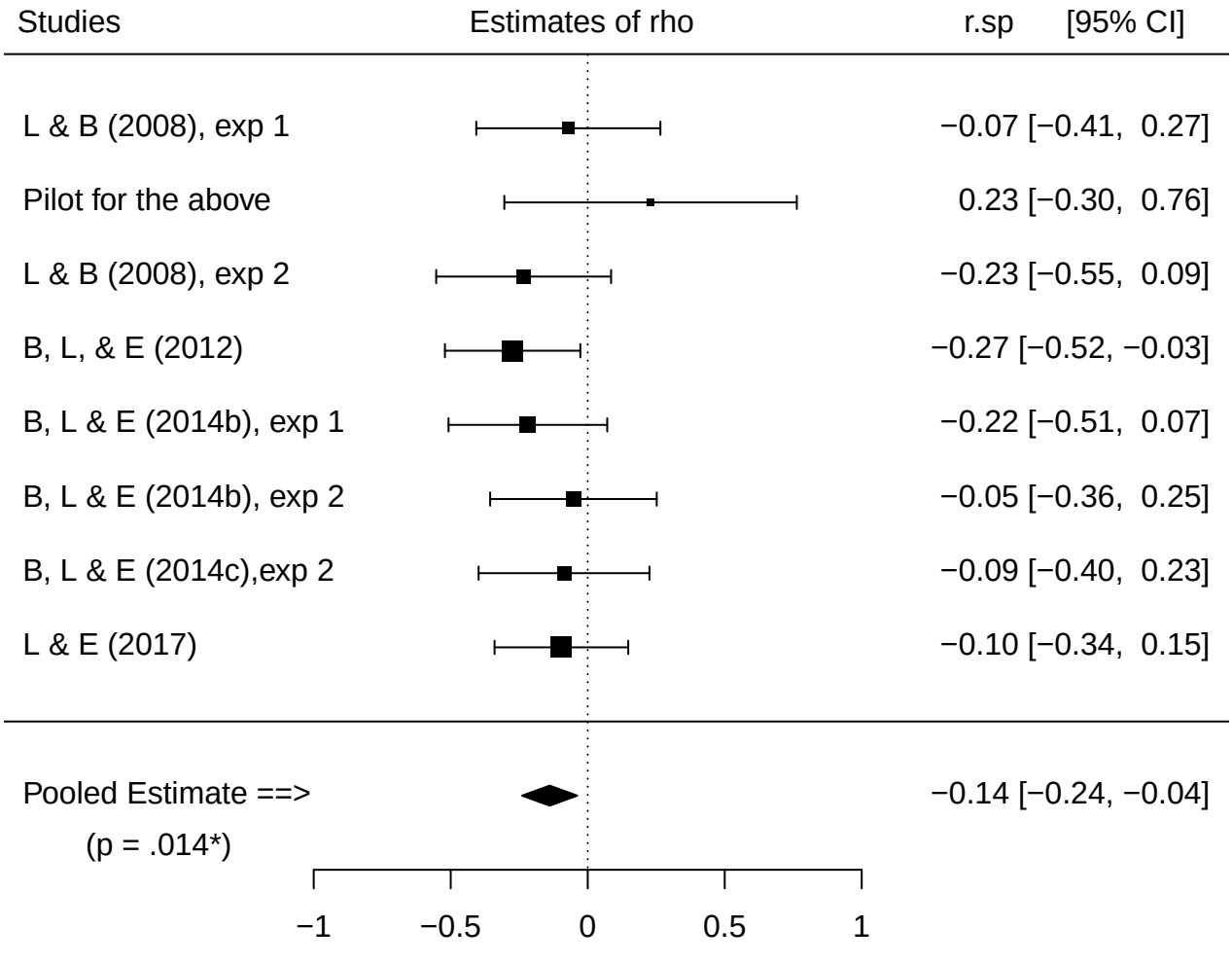
1	badly beaten	golf course	quick stop
2	badly injured	good guess	quick trip
3	bath soap	green grass	quiet corner
4	bean soup	grey hair	quiet room

5	bird bath	hair loss	raindrop*
6	bread bin	hard work	rapid rise
7	brick wall	heat loss	right hand
8	car crime	high price	ring road
9	cellar door	high rate	rocky slope
10	cheap seat	hill top	rubber ball
11	check list	hot drink	rubber glove
12	clear glass	important point	sadly misunderstood
13	cotton cloth	important thing	sadly neglected
14	cowboy*	job loss	safe place
15	current level	key hole	sandy shore
16	current pace	kind heart	sea salt
17	current trend	lamp light	sea side*
18	dark shape	land mass	sharp sound
19	daydream*	land use	short nap
20	deep hole	leather seat	slippery road
21	deep sea	life time	slippery slope
22	deep sleep	long life	small talk
23	designer jeans	long way	soft cloth
24	designer shirt	loud noise	soft ground
25	desk chair	loud sound	soft job
26	effective means	main gate	some chance
27	effective trick	main road	steam train
28	expensive deal	metal roof	stone steps
29	expensive piece	milkman*	strong bond
30	fair deal	name game	sudden halt
31	fair share	new car	sunlight*

32	fast food	nice place	sunset*
33	fine wine	nice try	tall man
34	firm hold	night light	tall tree
35	fish dish	no news	tea time
36	fish pond	note pad	time frame
37	floor lamp	paint brush	time span
38	free lunch	paper sack	tool box
39	free ride	phone call	town house
40	fresh air	plain talk	town square
41	fruit tree	plastic pipe	water mill
42	full force	popular appeal	west wind
43	full speed	popular demand	wild child
44	garden gate	price war	wish list
45	gas tank	private collection	wood frame
46	gift card	private property	workplace*
47	gift list		

Notes.

a. Asterisks highlight eight targeted expressions which that seem easy to interpret compositionally (i.e., as two words) but which are commonly spelled as one word in contemporary English.
b. The number of MWEs having two, three, four, and five syllables is, respectively: 108, 18, 13, and 1.
c. All but two of the MWEs (i.e., *no news* and *some chance*) have the phrase structure Adj-N or N-N.



https://itl.editorialmanager.com/

lindstromberg

https://itl.editorialmanager.com/l.asp?i=17330&l=PXYX3EAS

Reviewer #3:

…"I only have a few comments (mostly related to the structure of the text):

- Research questions: "with respect to each of our three measures": it would be useful to repeat the measures here (whole MWE frequencies, MWE total constituent word frequencies, frequencies of the least frequent word in each MWE).

This has been done.

- It would be easier to read the conclusion if it were more clearly organized in function of the three measures of frequency or the research questions.

A good deal of reorganization and rewriting has been done in the first part of the original conclusion response to this recommendation.

- The conclusion section is rather long. I think I would add separate sections: discussion, limitations, conclusion,…

This has been done.

Thank you!

Seth Lindstromberg