

# Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence

Dana Gablasova, Vaclav Brezina, and Tony McEnery

Lancaster University

This article focuses on the use of collocations in language learning research (LLR). Collocations, as units of formulaic language, are becoming prominent in our understanding of language learning and use; however, while the number of corpus-based LLR studies of collocations is growing, there is still a need for a deeper understanding of factors that play a role in establishing that two words in a corpus can be considered to be collocates. In this article we critically review both the application of measures used to identify collocability between words and the nature of the relationship between two collocates. Particular attention is paid to the comparison of collocability across different corpora representing different genres, registers, or modalities. Several issues involved in the interpretation of collocational patterns in the production of first language and second language users are also considered. Reflecting on the current practices in the field, further directions for collocation research are proposed.

**Keywords** corpus linguistics; collocations; association measures; second language acquisition; formulaic language

---

We wish to thank the anonymous reviewers and Professor Judit Kormos for their valuable comments on different drafts of this article. The research presented in this article was supported by the ESRC Centre for Corpus Approaches to Social Science, ESRC grant reference ES/K002155/1, and Trinity College London. Information about access to the data used in this article is provided in Appendix S1 in the Supporting Information online.

Correspondence concerning this article should be addressed to Dana Gablasova, Department of Linguistics and English Language, Lancaster University, Lancaster LA1 4YL, UK. E-mail: d.gablasova@lancaster.ac.uk

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

## Introduction

Formulaic language has occupied a prominent role in the study of language learning and use for several decades (Wray, 2013). Recently an even more notable increase in interest in the topic has led to an “explosion of activity” in the field (Wray, 2012, p. 23). Published research on formulaic language has cut across the fields of psycholinguistics, corpus linguistics, and language education with a variety of formulaic units identified (e.g., collocations, lexical bundles, collocations, collgrams) and linked to fluent and natural production associated with native speakers of the language (Ellis, 2002; Ellis, Simpson-Vlach, Römer, Brook O'Donnell, & Wulff, 2015; Erman, Forsberg Lundell, & Lewis, 2016; Howarth, 1998; Paquot & Granger, 2012; Powley & Syder, 1983; Schmitt, 2012; Sinclair, 1991; Wray, 2002). Language learning research (LLR) in both first and second language acquisition (SLA) has focused on examining the links between formulaic units and fundamental cognitive processes in language learning and use, such as storage, representation, and access to these units in mental lexicon (Ellis et al., 2015; Wray 2002, 2012, 2013). Recent studies provide compelling evidence that formulaic units play an important role in these processes and are psycholinguistically real (Ellis, Simpson-Vlach, & Maynard, 2008; Schmitt, 2012; Wray, 2012). Similarly, formulaic expressions have long been at the forefront of interest in corpus linguistics (CL). Corpora represent a rich source of information about the regularity, frequency, and distribution of formulaic patterns in language. In CL, particular attention has been paid both to techniques that can identify patterns of co-occurrence of linguistic items and to the description of these formulaic units as documented in language corpora (e.g., Evert, 2005; Gries, 2008; Sinclair, 1991). As demonstrated by a number of studies to date, combining data, methods, and models from LLR and CL has a significant potential for the investigation of language acquisition by both first language (L1) and second language (L2) speakers. Yet, as researchers involved in both fields stress repeatedly (e.g., Arrpe, Gilquin, Glynn, Hilpert, & Zeschel, 2010; Durrant & Siyanova-Chanturia, 2015; Gilquin & Gries, 2009; Wray, 2002, 2012), this fruitful cross-pollination cannot succeed without careful consideration of the methods and sources of evidence specific to each of the fields.

This article seeks to contribute to the productive cooperation between LLR and CL by focusing on collocations, a prominent area of formulaic language that is of interest to researchers in both LLR and CL. Collocations, as one of the units of formulaic language, have received considerable attention in corpus-based language learning studies in the last 10 years (e.g., Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Ellis, Simpson-Vlach, & Maynard, 2008; González Fernández & Schmitt, 2015; Nesselhauf, 2005; Nguyen

& Webb, 2016; Paquot & Granger, 2012). These studies have used corpus evidence to gain information about collocational patterns in the language production of L1 as well as L2 speakers; patterns identified in corpora have also been used to form hypotheses about language learning and processing that were then more directly studied by experimental methods (Durrant & Siyanova-Chanturia, 2015; Gilquin & Gries, 2009).

The corpus-based collocational measures used in these studies are of paramount importance as they directly and significantly affect the findings of these studies and consequently the insights into language learning that they provide. However, while efforts have been made to standardize the conflicting terminology that inevitably arises from such a large body of research (Ebeling & Hasselgård, 2015; Evert, 2005; Granger & Paquot, 2008; Nesselhauf, 2005; Wray, 2002), the rationale behind the selection of collocational measures in studies on formulaic development is not always fully transparent and systematic, making findings from different studies difficult to interpret (González Fernández & Schmitt, 2015). While considerable research in CL has been devoted to a thorough understanding of these measures, their effects, and the distinction between them (e.g., Bartsch & Evert, 2014), in LLR collocation measures have not yet received a similar level of attention. The main objective of this article is thus to bridge the work on collocations in these two disciplines; more specifically, the article seeks to contribute, from the corpus linguistics perspective, to corpus-based LLR on collocations by discussing how to meaningfully measure and interpret collocation use in L1 and L2 production, thereby making the findings more systematic, transparent, and replicable. The article argues that more consideration needs to be given to the selection of the appropriate measures for identifying collocations, with attention paid to the linguistic patterns that underlie these measures and the psycholinguistic properties these patterns may be linked to. Next, the article addresses the application of collocation measures to longer stretches of discourse, pointing to the need to reflect on the effect of genre/register on the collocations identified in different (sub)corpora. Finally, we revisit some of the findings on collocation use by L1 and L2 users obtained so far, demonstrating how understanding of collocation measures can help explain the trends observed in language production. A clearer grasp of the key concepts underlying the study of collocations explained in this article will in turn lead to the creation of a theoretically and methodologically sound basis for the use of corpus evidence in studies of formulaicity in language learning and use.

Different approaches to operationalizing the complex notion of formulaicity and classifying collocations have been noted in the literature

(McEnery & Hardie, 2011, pp. 122–133). The two most distinct approaches typically recognized are the “phraseological approach,” which focuses on establishing the semantic relationship between two (or more) words and the degree of noncompositionality of their meaning, and the “distributional” or “frequency-based” approach, which draws on quantitative evidence about word co-occurrence in corpora (Granger & Paquot, 2008; Nesselhauf, 2005; Paquot & Granger, 2012). Within the latter approach, which is the focus of this article, three subtypes—surface, textual, and syntactic co-occurrences—can be distinguished (Evert, 2008) depending on the focus of the investigation and the type of information used for identifying collocations. While the surface co-occurrence looks at simple co-occurrence of words, textual and syntactic types of co-occurrence require additional information about textual structure (e.g., utterance and sentence boundaries) and syntactic relationships (e.g., verb + object, premodifying adj. + noun).

When identifying collocations, we also need to consider the distance between the co-occurring words and the desired compactness (proximity) of the units. Here, we can distinguish three approaches based on n-grams (including clusters, lexical bundles, congrams, collgrams, and p-frames), collocation windows, and collocation networks. The n-gram approach identifies adjacent combinations such as *of the, minor changes*, and *I think* (these examples are called bigrams, i.e., combinations of two words) or adjacent combinations with possible internal variation such as *minor but important/significant/observable changes*. The collocation window approach looks for co-occurrences within a specified window span such as *5L 5R* (i.e., five words to the left of the word of interest and five to the right), thus identifying looser word associations than the n-gram approach. Using the window approach, the collocations of *changes* will include *minor* (as in the n-gram approach) but also *notify* or *place*, showing broader patterns and associations such as *notify somebody of any changes* and *changes which took place* (all examples are from the British National Corpus [BNC]). Finally, collocation networks (Brezina, McEnery, & Wattam, 2015; Phillips, 1985; Williams, 1998) combine multiple associations identified using the window approach to bring together interconnections between words that exist in language and discourse. The distance between the node and the second-, third-, and so on order collocate is thus not the immediate window distance (as is the case for first-order collocates), but a mediated distance via an association with another word in the network.

To illustrate the issues in corpus-based LLR, this study will use surface-level collocations and the window approach (which includes bigrams as a special case of the window approach). This is the broadest approach, often followed in

corpus linguistics (McEnery & Hardie, 2011), which can be further adjusted if the research requires identification of a particular type of linguistic structures.

## Identifying Collocational Patterns in Corpora

### Operationalizing Formulaicity: Linking Corpus Evidence to Psycholinguistic Reality

Corpora can provide direct information about the formulaic usage patterns produced by L1 and L2 users; they are also an indirect, approximate source of information about experience with (exposure to) language use that plays a role in the cognitive processing and representation of language (e.g., Ellis, 2014; Ellis et al., 2015; Rebuschat & Williams, 2012). **A selection of a particular collocational measure should always be preceded and motivated by a reflection on the dynamics of language production and its connection to psycholinguistic processes.** Language learning studies using corpus-based statistical definitions of collocations have traditionally distinguished between two major criteria of collocability (Ellis et al., 2015; Schmitt, 2012), absolute frequency and strength of association between word combinations. While the frequency-only approach relies solely on counting the co-occurrences of word forms, association measures (AMs) combine information about frequency with other collocational properties that can be expressed mathematically (e.g., Evert, 2005, 2008; Hunston, 2002; McEnery & Hardie, 2011). Although AMs are often grouped together and referred to as measures of strength of word combinations this may unhelpfully collate a range of collocational properties that should be, if possible, isolated or, if not, acknowledged because they play different roles in language processing. To illustrate this, the following discussion explores the frequency aspect of collocation and three dimensions of formulaicity related to frequency: dispersion, exclusivity, and directionality.

The frequency of linguistic structures is undoubtedly a major variable in LLR, with strong links to psycholinguistic processes involved in language learning such as noticing, representation, access, and production of language (e.g., Ellis, 2002, 2014; Rebuschat & Williams, 2012). With respect to the acquisition and production of formulaic language, both L1 and L2 speakers have been found to show sensitivity to the frequency-based distribution of word combinations (e.g., Ellis, 2002; Ellis et al., 2015; González Fernández & Schmitt, 2015; Sosa & MacFarlane, 2002; Wray, 2002). However, the causal relationship between frequency and collocational knowledge is not straightforward and frequency-only definitions of formulaicity may “collapse distinctions that intuition would deem relevant” (Simpson-Vlach & Ellis, 2010, p. 488) with other cognitive predictors of language learning being a factor (such as proficiency of

learners and salience, uniqueness, semantic meaning, or personal relevance of expressions) (e.g., Gass & Mackey, 2002; Wray, 2012).

Raw (absolute) frequency, which has so far been used in corpus-based LLR, while a good measure of overall repetition in language, may not be the best predictor of the regularity and predictability in language use. Corpus findings show that sometimes even fairly frequent co-occurrences of words appear only in a very particular context in language or are produced by a very small number of speakers/writers. For example, in the BNC, *risk issues* and *moral issues* have very similar absolute frequency (54 and 51 occurrences, respectively). However, while all 54 instances of the first expression occurred in one text sample, the latter occurred in over 41 texts. This distributional pattern changes the probability of the collocation occurrence in language and the likelihood of any speakers' experience with or activation of such a unit. To obtain a fuller picture of the role of frequency in collocational knowledge and use, the dispersion of collocations in a corpus should also be considered (Brezina et al., 2015; Gries, 2010, 2013). The *dispersion* of a linguistic feature expresses how (un)evenly this feature occurs across the corpus and can be used as a proxy measure of occurrence regularity. Dispersion is thus an important predictor in language learning because collocations that are more general (i.e., occur across a variety of contexts) are more likely to be encountered by language users regardless of the context of use.

The second dimension to be discussed is the *exclusivity* of collocates, that is, the extent to which the two words appear solely or predominantly in each other's company, usually expressed in terms of the relationship between the number of times when they are seen together as opposed to the number of times when they are seen separately in the corpus (e.g., the Mutual Information [MI] score highlights this property). Exclusivity is likely to be strongly linked to predictability of co-occurrence, when the appearance of one part of the collocation brings to mind the other part. For example, collocations such as *zig zag*, *okey dokey*, and *annus mirrabilis* are fairly exclusively associated. We could hypothesize that words that are likely to be seen in each other's company may be more easily recognized, acquired, and stored as a unit. We could also expect stronger priming effects between the two words (Wray, 2002). Finally, the degree of exclusivity could be positively correlated with salience and hence noticing (e.g., Gass & Mackey, 2002).

The third dimension to be considered in the processing and learning of collocations is *directionality* (Brezina et al., 2015; Gries, 2013; Handl, 2008). Directionality is a concept that postulates that the components in a collocation do not attract each other with equal strength (i.e., the attractions are often

asymmetrical); in other words, each of the two words in the collocation involves different degree of probability that it will occur with the other word in the pair. Thus we may be able to predict one word on the basis of the other word in the pair but not the other way around. For example, while *decorations* may prime speakers for *Christmas* (in the BNC 11% of instances of *decorations* are preceded by *Christmas*), this would not work with the same strength if we were shown *Christmas* (only 0.5% of instances of *Christmas* are followed by *decorations*); another example may be *extenuating circumstances* with *extenuating* priming *circumstances* in a much stronger way than vice versa. This dimension of collocation may be relevant to studies on mental representation that involve priming or completion tasks—one of the words may prime more strongly than the other. The AM that grasps this dimension is Delta P (Gries, 2013).

This section highlighted three major dimensions of word collocability with likely effects on processing of linguistic evidence and development of collocational knowledge by learners. These properties should be considered before selecting a method or AM for identifying (operationalizing) collocations in language as each of these will highlight different aspects of formulaicity between two words.

### Selecting and Interpreting AMs

The previous section discussed three collocational properties in general terms. This section will focus on specific AMs and demonstrate the level of understanding necessary before we select specific AMs and interpret findings based on them. Despite the existence of dozens of AMs, so far only a limited set has been used in research, with the t-score and MI-score holding a dominant position in most recent studies (e.g., Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Ellis et al., 2008; Granger & Bestgen, 2014; Nguyen & Webb, 2016; Siyanova-Chanturia, 2015). Unfortunately, so far, the statistical AMs in LLR have been largely used as apparently effective, but not fully understood mathematical procedures. As González Fernández and Schmitt (2015, p. 96) note, “it is not clear which of these [MI-score and t-score] (or other) measures is the best to use in research, and to date, the selection of one or another seems to be somewhat arbitrary.” Consequently, we will discuss three specific AMs, t-score, MI-score, and Log Dice, and consider their ability to highlight different aspects of formulaicity. The t-score and MI-score were chosen because of their prominent role in recent corpus-based studies; Log Dice is introduced as an alternative to the MI-score. For the proper (informed) use of each of the AMs, we need to understand (1) the mathematical reasoning behind the measure, (2) the scale on which it operates, and (3) its practical effect (what combinations

of words get highlighted and what gets hidden/downgraded). A full mathematical justification of the claims below (and further examples) can be found in Appendix S2 in the Supporting Information online. Let us now proceed to critically explore some of the major scores with regard to the three features outlined.

### *T-score*

The t-score has been variously labeled as a measure of “certainty of collocation” (Hunston, 2002, p. 73) and of “the strength of co-occurrence,” which “tests the null hypothesis” (Wolter & Gyllstad, 2011, p. 436). These conceptualizations are neither particularly helpful nor accurate. As Evert (2005, pp. 82–83) shows, although originally intended as a derivation of the *t* test, the t-score does not have a very transparent mathematical grounding. It is therefore not possible to reliably establish the rejection region for the null hypothesis (i.e., statistically valid cutoff points) and interpret the score other than as a broad indication of certain aspects of the co-occurrence relationship (see below). The t-score is calculated as an adjusted value of collocation frequency based on the raw frequency from which random co-occurrence frequency is subtracted. This is then divided by the square root of the raw frequency. Leaving aside the problematic assumption of the random co-occurrence baseline (see “MI-score” below) the main problem with the t-score is connected with the fact that it does not operate on a standardized scale and therefore cannot be used to directly compare collocations in different corpora (Hunston, 2002) or to set reliable cutoff point values for the results. At this point, a note needs to be made about the different levels of standardization of measurement scales; this discussion is also relevant for the MI-score and Log Dice (see below). The most basic level involves no standardization. For example, raw frequency counts or t-scores are directly dependent on the corpus size, that is, they operate on different scales, and are thus not comparable across corpora of different sizes. The second, more advanced, level involves normalization, which means an adjustment of values to one common scale, so that values from different corpora are directly comparable. For example, percentages or relative frequencies per million words operate on normalized scales. Finally, the most complex level is based on scaling of values, which involves a transformation of values to a scale with a given range of values. For example, the correlation coefficient (*r*) operates on a scale from -1 to 1.

In practice, as has been observed many times (e.g., Durrant & Schmitt, 2009; Hunston, 2002; Siyanova & Schmitt, 2008), the t-score highlights frequent combinations of words. Researchers also stress the close links between the t-score and raw frequency, pointing out that t-score rankings “are very similar to rankings based on raw frequency” (Durrant & Schmitt, 2009,



p. 167). This is true to some extent, especially when looking at the top ranks of collocates. For example, for the top 100 t-score-ordered bigrams in the BNC, the t-score strongly correlates with their frequency ( $r = 0.7$ ); however, the correlation is much weaker ( $r = 0.2$ ) in the top 10,000 bigrams. While the literature has stressed similarity between collocations identified with t-score and raw frequency, a less well understood aspect of t-score collocations is the downgrading (and thus effectively hiding) of word combinations whose constituent words appear frequently outside of the combination. For instance, the bigrams that get downgraded most by t-score ranking in the BNC are *is the*, *to a*, and *and a*, while combinations such as *of the*, *in the*, and *on the* retain their high rank on the collocation list despite the fact that both groups of bigrams have a large overall frequency. The t-score and frequency thus cannot be seen as co-extensional terms as suggested in the literature. Instead the logic of their relationship is this: While all collocations identified by the t-score are frequent, not all frequent word combinations have a high t-score.

### *MI-score*

The MI-score has enjoyed a growing popularity in corpus-based LLR. It is usually described as a measure of strength (e.g., Hunston, 2002) related to tightness (González Fernández & Schmitt, 2015), coherence (Ellis et al., 2008), and appropriateness (Siyanova & Schmitt, 2008) of word combinations. It has also been observed to favour low-frequency collocations (e.g., Bestgen & Granger, 2014) and it has been contrasted with the t-score as a measure of high-frequency collocations, although this dichotomous description is too general to be useful in LLR. The MI-score uses a logarithmic scale to express the ratio between the frequency of the collocation and the frequency of random co-occurrence of the two words in the combination (Church & Hanks, 1990)—the random co-occurrence is analogous to the corpus being a box in which we have all words written on separate small cards—this box is then shaken thoroughly. Whether this model of random occurrence of words in a language is a reliable baseline for the identification of collocations is questionable, however (e.g., Stubbs, 2001, pp. 73–74). In terms of the scale, the MI-score is a normalized score that is comparable across language corpora (Hunston, 2002), although it operates on a scale that does not have a theoretical minimum and maximum, that is, it is not scaled to a particular range of values. The value is larger the more exclusively the two words are associated and the rarer the combination is. We must therefore be careful not to automatically interpret larger values, as has been done often (see above), as signs of stronger, tighter, or more coherent word combinations, because the MI-score is not constructed as a (reliable) scale for

coherence or semantic unity of word combinations; coherence and semantic unity are an indirect side effect of the measure's focus on rare exclusivity.

Highlighting rare exclusivity is thus the main practical effect of the mathematical expression of the MI-score. It is therefore somewhat misleading to claim that the "[MI-score] is not so strongly linked with frequency as other association measures" (Siyanova & Schmitt, 2008, p. 435). On the contrary, the MI-score is negatively linked to frequency in that it rewards lower frequency combinations, for which less evidence exists in the corpus. For example, the combination *ceteris paribus* (*freq.* = 46, *MI* = 21) receives a lower score than the name *jampa ngodrup* (*freq.* = 10, *MI* = 23.2), although both are exclusively associated (i.e., the constituent words occur in the BNC only in these combinations) and the former combination is almost five times as frequent as the latter. The low-frequency bias of the MI-score is fixed in MI2 (where the collocation frequency is squared), a version of the MI-score that does not penalize frequency and awards *jampa ngodrup* and *ceteris paribus* the same score (*MI*2 = 26.5). Unfortunately, MI2 has not yet received any attention in LLR. Overall, the MI-score strongly favors names (if proper nouns are considered); terms; and specialized or technical, low-frequency combinations (e.g., *carbonic anhydrase*, *yom kippur*, *afrika korps*, *okey dokey*<sup>1</sup>) and thus highlights collocations that are not equally distributed across language, precisely because low frequency items are often restricted to specific texts or genres.

### Log Dice

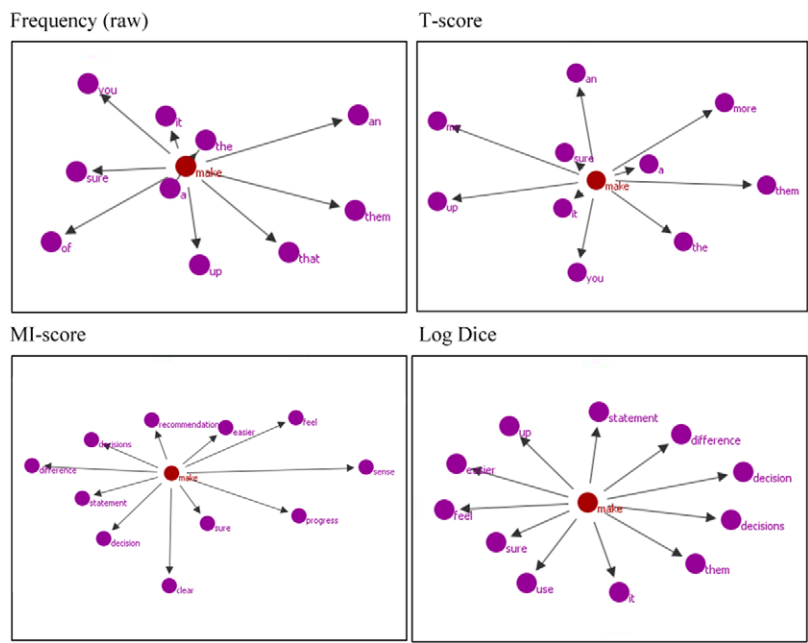
Log Dice is a measure that has not yet been explored in LLR. Log Dice takes the harmonic mean (a type of average appropriate for ratios) of two proportions that express the tendency of two words to co-occur relative to the frequency of these words in the corpus (Evert, 2008; Smadja, McKeown, & Hatzivassiloglou, 1996). Log Dice is a standardized measure operating on a scale with a fixed maximum value of 14, which makes Log Dice directly comparable across different corpora and somewhat preferable to the MI-score and MI2, neither of which have a fixed maximum value. With Log Dice, we can thus see more clearly than with MI or MI2 how far the value for a particular combination is from the theoretical maximum, which marks an entirely exclusive combination.

In its practical effects, Log Dice is a measure fairly similar to the MI-score (and especially to MI2) because it highlights exclusive but not necessarily rare combinations (the latter are highlighted by the original version of the MI-score). Combinations with a high Log Dice (over 13) include *femme fatale*, *zig zag*, and *coca cola*, as well as the combinations mentioned as examples with a high MI-score. For 10,000 Log-Dice-ranked BNC bigrams, the Pearson's

correlations between Log Dice and MI and Log Dice and MI2 are 0.79 and 0.88, respectively, showing a high degree of similarity, especially between Log Dice and MI2; by comparison, the correlation between Log Dice and the t-score is only 0.29. Like the MI-score and MI2, Log Dice can be used for term-extraction, hence the description of Log Dice as “a lexicographer-friendly association score” (Rychlý, 2008, p. 6). Unlike the MI-score, MI2, and the t-score, Log Dice does not invoke the potentially problematic shake-the-box, random distribution model of language because it does not include the expected frequency in its equation (see Appendix S2 in the Supporting Information online). More importantly, Log Dice is preferable to the MI-score if the LLR construct requires highlighting exclusivity between words in the collocation with a clearly delimited scale and without the low-frequency bias.

In sum, we have discussed the key principles of three main and one related AMs and the practical effects that these measures have on the types of word combinations that get highlighted/downgraded. It is important to realize that AMs provide a specific system of collocation ranking that differs from raw frequency ranking, prioritizing aspects such as adjusted frequency (t-score), rare exclusivity (MI), and exclusivity (MI2, Log Dice). As a visual summary, Figure 1 displays the differences between raw frequency and the three main AMs (t-score, MI-score, and Log Dice) in four simple collocation graphs for the verb *to make* in a one-million-word corpus of British writing (British English 2006). In these graphs, we can see not only the top 10 collocates identified by each metric, but also the strength of each association—the closer the collocate is to the node (*make*), the stronger the association. In addition, we can and should explore alternative AMs that capture other dimensions of the collocational relationship such as directionality (Delta P) and dispersion (Cohen's d); due to space constraints, these are not discussed here. As a general principle in LLR, however, we should critically evaluate the contribution of each AM and should not be content with one default option, no matter how popular.

A possible reason for the relatively narrow range of AMs in general use is that until recently it was difficult to calculate different AMs because the majority of corpus linguistic software tools supported only a very limited AM range; this might partly explain the popularity of the MI-score and the t-score and the underexploration of other measures. GraphColl (Brezina et al., 2015), the tool used to produce the collocations and their visualisations in Figure 1, was developed with the specific aim of allowing users to easily apply dozens of different AMs while supporting high transparency and replicability through explicit access to the equation used in each AM. In addition to existing AMs,



**Figure 1** Top 10 collocations of *make* for frequency and three AMs using L0, R2 windows in the BE06 corpus. [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

it allows users to define their own collocational measure. Another possibility for calculating a broad range of AMs is to use the statistical package R (R Core team, 2016), which is becoming increasingly popular in LLR; however, unlike GraphColl, R requires the analyst to have experience with coding and command-line operations.

The first part of this article addressed a range of theoretical concepts and statistical measures related to collocations used in corpus-based LLR; the following section focuses on applying these concepts and measures to larger stretches of language in corpora in order to examine the degree and nature of formulaicity in the production of L1 and L2 speakers and possible sources of variation in their collocational use.

**Comparing Collocations Across Different Linguistic Settings:  
Stability of Association Strength**

The aim of this section is to examine to what extent the strength of association between two words varies according to linguistic settings (e.g., involving

different social and situational characteristics) as represented in different (sub)corpora. The issue of the effect of linguistic setting on the strength of word association is crucial for the further development of a model of formulaicity in language use (Ellis et al., 2015). Ample evidence from across different dimensions of language use (e.g., morpho-syntactic, pragmatic, and lexical) demonstrates that linguistic setting is a major predictor of variation in language (e.g., Biber et al., 1999). A systematic, context-based variation has also been observed with respect to different aspects of formulaicity, with substantial attention given to collocations typical of the academic register (e.g., Howarth, 1998; Simpson-Vlach & Ellis, 2010). The studies so far have focused on identifying formulaic units typical of a particular register or genre and described the frequency of occurrence of these expressions according to these settings (e.g., Ackermann & Chen, 2013); the strength of collocations, a key factor in numerous studies on formulaicity, has, however, not been examined. As a result, little empirical evidence is available about the degree to which genres, registers, and modes of communication affect the strength of association between words.

The variation in collocational strength across linguistic settings also has considerable methodological implications for corpus-based LLR. Given the evidence for the effect of linguistic setting on frequency of specific words (Biber et al., 1999), variation in collocational strength due to linguistic setting should also be expected, because frequency is a key piece of information in the calculation of association strength. Despite this indirect evidence, numerous studies establish collocational strength of word combinations on general corpora that include a variety of registers and/or genres as well as spoken and written modalities without critically engaging with the source of collocational strength in these corpora. For example, several recent studies used the whole BNC or the whole Corpus of Contemporary American English, both large corpora consisting of multiple genres/register, as the basis for establishing collocational strength in native-speaker production (Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Granger & Bestgen, 2014; Siyanova & Schmitt, 2008). Moreover, some studies establish collocational strength using corpora containing one specific genre (e.g., journalistic writing as in Durrant & Schmitt, 2009, and Siyanova-Chanturia, 2015) to evaluate collocations in a different genre (e.g., L2 academic writing). While use of these reference corpora may be necessary in the absence of a more suitable comparable corpus, the possible limitations of such a comparison need to be examined and acknowledged.

To illustrate the impact of different genres, registers, and modality on collocational strength and the need for further research in this area, we compared

**Table 1** Overview of (sub)corpora used

Corpus	Size	Representativeness
British National Corpus (BNC)	98,560,118	Written and spoken (10M), different registers
BNC_A	15,778,043	Written, academic writing
BNC_N	9,412,245	Written, news
BNC_F	16,143,913	Written, fiction
BNC – Context governed (BNC_CG)	6,196,134	Spoken, formal
BNC – Demographic (BNC_D)	4,234,093	Spoken, informal
BNC – 2014 Spoken (BNC_SP)	4,789,185	Spoken, informal
CANCODE (CANC)	5,076,313	Spoken, informal

collocational strength in different corpora and subcorpora representing a variety of modes, registers, and genres in British English. First, we used the BNC as a whole and three of its written subcorpora (academic writing, newspapers, and fiction) and two spoken subcorpora (formal and informal speech). Second, CANCODE, a five-million-word corpus representing informal British speech in the late 1990s was used. Finally, a five-million-word sample from the spoken section of the BNC 2014 (BNC\_SP) was used. Table 1 provides an overview of the (sub)corpora used. A more detailed description of each corpus as well as information about how to access it can be found in Appendix S1 in the Supporting Information online. The whole BNC was included as it has traditionally been used in collocational research as a reference corpus (e.g., Durrant & Schmitt, 2009; Granger & Bestgen, 2014). Five BNC subcorpora were used to investigate the variation in collocational strength inside the BNC. CANCODE and BNC\_SP were selected to strengthen the spoken component of the study by looking at more recent corpora of informal speech that are directly comparable with the informal subcorpus (i.e., BNC-Demographic) from the BNC.

We selected three types of collocations representing a range of constructions that commonly appear in collocational research (Durrant & Schmitt, 2009; Ebeling & Hasselgård, 2015; Granger & Bestgen, 2014; Paquot & Granger, 2012; Siyanova & Schmitt, 2008): verb + complementation (*make + sure/decision/point*), adjective + noun (*human + beings/rights/nature*), and adverb + adjective (*vitality/very/really + important*). Due to space constraints, only the results for the first construction, that is, verb + complementation, are presented in Table 2; results for the other two structures, which echo the points made here, can be found in Appendix S3 in the Supporting Information online. To address different aspects of formulaicity, three AMs discussed in the

**Table 2** *Make* (lemma), [L0 R2]

	BNC	BNC_A	BNC_N	BNC_F	BNC.CG	BNC.D	BNC.SP	CANC
<i>MI-score</i>								
sure	6.80	7.09	7.26	5.78	6.90	6.64	6.26	6.92
decision	4.55	3.67	4.07	5.86	6.12	7.91	7.57	8.07
point	3.44	2.92	3.84	3.68	4.11	3.12	3.01	3.93
<i>t-score</i>								
sure	74.52	13.58	21.45	32.24	28.59	16.92	18.36	22.31
decision	27.59	9.44	9.22	10.45	11.58	5.08	6.82	10.82
point	27.41	9.50	7.78	35.61	13.78	3.43	4.20	6.47
<i>Log Dice</i>								
sure	9.63	7.60	9.52	9.61	10.68	10.17	10.07	10.61
decision	6.91	6.61	7.16	6.62	8.31	7.05	7.60	8.87
point	6.90	6.65	6.74	6.55	8.52	6.03	6.30	7.30

previous section were used. The full frequency table on which the calculation of AMs is based is also available in Appendix S3 in the Supporting Information online.

Considering the results, the t-score, as expected, varies greatly across the different corpora, demonstrating its strong dependence on corpus size. Corpus comparisons based on this measure thus cannot be made easily as the t-score will be larger or smaller depending upon the corpus size regardless of the strength of the collocation. While the differences in t-scores between the three similarly sized corpora of informal speech (BNC.D, BNC.SP, and CANC) are comparatively smaller, they are still difficult to interpret due to the absence of a commensurable scale. By contrast, the overall variation across the MI-score and Log Dice values is considerably smaller, demonstrating their independence from corpus size, which makes them more directly comparable across corpora as discussed earlier (see also Hunston, 2002). For those reasons, the following discussion will focus on the findings based on these two measures only.

The variation in the collocational strength between the BNC and its various subcorpora suggests that large aggregate data such as the whole BNC may hide different distributions of formulaicity across registers and genres as well as across the written/spoken divide. Although this is noticeable for each

collocation, it is more striking in some cases. While, for example, for *make sure* the variation in the MI-score values is relatively small, for *make [a] decision* the MI-score values range from 3.67 to 7.91 and the Log Dice values from 6.61 to 8.31. As shown in Appendix S3 in the Supporting Information online, these differences are not uncommon (e.g., the range of MI-score values for *human nature* is 6.48 to 10.92; the range of Log Dice for *vitality important* is 5.00 to 8.18). It is also worth noting that the order of collocations according to the strength of association varies between subcorpora. For example, *make sure* is more strongly associated on the MI-score in the whole BNC and in the academic and newspaper subcorpora (BNC\_A and BNC\_N) than in the corpora of informal speech (BNC\_D and BNC\_SP) while the opposite is true of *make [a] decision*.

When looking at the three corpora representing informal spoken English (BNC\_D, BNC\_SP, and CANC) the variation is considerably smaller for both the MI-score and Log Dice, although their values still differ by more than one unit (e.g., Log Dice for *make [a] point* ranges from 6.03 to 7.3 and *make [a] decision* from 7.05 to 8.87). These illustrative results demonstrate that the relationship between formulaicity and language produced in different settings needs careful attention.

These findings have practical implications for interpreting the data from corpus-based studies. Variation across the MI-score and Log Dice in corpora representing the same type of discourse (as illustrated by the three corpora of spoken, informal English) suggests that very fine-grained categorization of AM values may not be entirely valid. For example, Durrant and Schmitt (2009) initially categorize MI-scores by one unit (e.g., 3.99 to 4.99). A more general categorization into lower and higher values appears more meaningful (Granger & Bestgen, 2014). Moreover, given the possibility of variation as an artifact of the corpus design, it is advised that the data are obtained from several corpora before being used as independent measures or predictors in LLR. However, further research is needed to provide valid guidelines in this area.

Systematic variation in collocational patterns (e.g., their strength and frequency) across genres, registers, and modes, as suggested by our exploration, has implications for research that seeks to shed light on the links between linguistic experience (exposure) of speakers and their collocational knowledge and use. For example, in the case of L2 acquisition, research suggests that even the L2 speakers living in the country of the target language may have an imbalanced or limited exposure to different domains (e.g., genres) of language with implications for language proficiency, dominance, and activation in these domains (Grosjean, 2010). This may affect their exposure to the word



combinations occurring in these domains and their characteristics (e.g., the strength with which they collocate).

### **Interpreting the Evidence: Collocational Usage Patterns in L1 and L2 Production**

Understanding the processes behind the co-selection of words has been one of the core issues in research on formulaicity (e.g., Arrpe et al., 2010; Nesselhauf, 2005; Siyanova-Chanturia & Martinez, 2015). Corpora, as large databases that document the products of users' word selection and co-selection, can reveal regularities in collocational preferences of users, allowing researchers to hypothesize about the factors involved in the acquisition, representation, and production of these word combinations. The aim of this section is to demonstrate that interpretation of corpus evidence is closely related to the full understanding of the practical effects of individual AMs.

To highlight issues in current LLR on collocations as well as areas for future inquiry, we will revisit one of the key findings from corpus-based SLA studies on collocational use of L2 speakers. These studies compared L1 and L2 users to investigate similarities/differences in the co-selection of words by these two groups and repeatedly reported that L1 users systematically produced collocations with higher MI-score values than L2 users (e.g., Durrant & Schmitt, 2009; Ellis et al., 2015; Granger & Bestgen, 2014; Schmitt, 2012). This difference has led Schmitt (2012, p. 6) to conclude that "the lack of these 'MI' collocations is one key feature which distinguishes native from nonnative production." As the source of L1/L2 difference in these studies is closely tied to a specific AM (the MI-score), this finding represents a good opportunity for a discussion of the effects of a particular measure on the understanding of collocations in LLR. Moreover, given its growing prominence and popularity across various areas of LLR, the MI-score deserves further exploration, although on a more general level the issues raised here are applicable to any AM. The strength of collocations produced by L1 and L2 speakers in these studies (e.g., Durrant & Schmitt, 2009) is established by extracting all target word combinations (e.g., all adjective + noun combinations) from the L1 and L2 corpora. Each of these combinations is then checked in a reference corpus (usually a large, general corpus of the target language such as the BNC for English) and its MI-score (or other relevant AM score) in this corpus is calculated. Next, researchers look at the combinations found in the L1 and L2 production and compare them in terms of the scores (collocational strength) established on the basis of the reference corpus. Based on this procedure, researchers can see the proportion of word combinations in L2 production with high MI-score values and compare this with the L1 speakers.

### Collocational Knowledge and Vocabulary Knowledge

As mentioned above, one of the major differences between L1 and L2 collocational production was related to the proportion of collocations with high MI-score values (the values were established on the basis of reference corpora such as the BNC). Compared to L1 users, collocations found in L2 production were more likely to appear frequently in the reference corpora and were less strongly associated on the MI-score; by contrast, L1 users produced more collocations with higher MI-score values (e.g., Ebeling & Hasselgård, 2015; Granger & Bestgen, 2014). As discussed in the section “MI-score,” the strength in this AM is a result of the interaction of exclusivity and the low frequency of the word co-occurrence. While the exclusivity aspect of the combinations is likely to play a role in the L1/L2 difference, equal attention should be given to the effects of the frequency of the word combination and its effect on the nature of collocational knowledge and selection by L1 and L2 users.

As pointed out, the MI-score tends to highlight infrequent combinations, whose constituents “may also be infrequent themselves” (Schmitt, 2012, p. 6) and the collocations may be “more specialised” (Ebeling & Hasselgård, 2015, p. 211). These less frequent and more specialized lexical items, which the MI-score favors, may not be in the lexicon of L2 speakers yet. It is thus possible that by using the MI-score researchers do not only measure collocational knowledge (preferences in word combinations), but also lexical knowledge of infrequent lexical items. For example, when considering phrases such as *densely populated*, which have been repeatedly used to illustrate the type of collocations that L2 users “take longer to acquire” (Durrant & Schmitt, 2009, p. 174), we should carefully examine whether absence of these combinations in L2 production is linked to some aspect of collocational knowledge, that is, to speakers’ ability to connect words that are in their lexicons (e.g., they know both *populated* and *densely* but choose not to combine them) or whether the absence/low frequency of these combinations rather suggests that one or both of the items are not present in L2 speakers’ lexicon. To illustrate the point, the word *populated* was found to occur with a relatively low frequency (only seven times) in the International Corpus of Learner English (ICLE), a three-million-word corpus of essays written by L2 university-level students used in several corpus-based SLA studies (for comparison, the frequency of *populated* in the written component of the BNC is 3.66 per million words). When it occurred, *populated* was premodified four times by *densely* and once by *sparsely*, suggesting that if the L2 writers used *populated* they tended to use it with a strongly associated premodifier. However, the use of words such as *populated* (i.e., formal, academic, and even specialised) may be overall lower in L2 production especially

at lower proficiency levels (Zareva, Schwanenflugel, & Nikolova, 2005). As a result, when calculating the proportion of word combinations with high MI-score values in L1 and L2 production following the method described above, the finding of fewer of these combinations in L2 production may be indirectly reflecting the difference in the knowledge of low-frequency words rather than ability to produce words that are strongly associated.

The difference between lexical and collocational knowledge has implications for the nature of processes behind word selection and deserves further attention in research on collocational choices in L1 and L2. First, more attention should be given to the investigation of the relationship between collocational and single-word knowledge and development; research in this area is still scant (for a recent exception, see Nguyen & Webb, 2016). A better understanding of the interaction between the two dimensions of vocabulary knowledge would also allow for a better modeling of language acquisition. Second, to gain a more fine-grained understanding of the roles of frequency and exclusivity in users' co-selection of words, measures such as MI2 or Log Dice (which do not penalise high-frequency collocates) should be considered in future studies.

### **Linguistic Properties of Collocations: Combining Phraseological and Frequency-Based Approaches**

While the phraseological approach distinguishes different types of collocations with respect to their semantic unity and fixedness of form, frequency-based accounts are largely blind to these properties deriving the strength of association between words directly from the evidence about word co-occurrence (Granger & Paquot, 2008; Nesselhauf, 2005). Although this objective nature of collocations identified by AMs has been considered as one of the advantages of the approach (Nguyen & Webb, 2016), a subsequent structural and semantic analysis of collocations identified by high or low AM scores can be crucial for getting a clearer picture of collocational patterns associated with certain groups of speakers or a certain type of word co-selection.

A particularly important distinction for psycholinguistic collocational research, which is not yet commonly made in corpus-based SLA studies on collocations, is that between fully fixed units found in dictionary entries (e.g., compounds such as *guinea pig* or *vacuum cleaner*) and units with some freedom (choice) in their combinability (e.g., *very/highly/extremely* + *important*; Herbst, 2011). As pointed out above (section "MI-score") higher MI-score values tend to favor units with a high level of semantic and structural unity; words with less stringent restrictions on their combination will have lower MI-scores. For example, the majority of collocations with high MI-score values identified in

learner writing in Bestgen and Granger are “compound-like units” (2014, p. 34), with combinations such as *ping pong*, *ha ha*, *rocket launchers*, *vacuum cleaner*, *alcoholic beverage*, *ozone layer*, *fire extinguisher*, *korean peninsula*, *toxic substance*, *grand rapids*, and *ice cream* forming the top ten units. Compounds and compound-like units are likely to differ from the less semantically and structurally restricted collocations in terms of how much choice (optionality) is involved in their co-selection in users’ production and we can speculate that they are also likely to differ in terms of their acquisition and representation. An MI-score alone thus provides a one-sided picture of collocational knowledge, which needs to be carefully evaluated and complemented by other accounts.

### Lexical Properties of Collocation: The Effect of Topic

As already discussed above (see “Comparing Collocations Across Different Language Settings”), researchers need to be mindful of the effects of genres, registers, and modes on the strength of collocations regardless of the AMs used. In addition to these, attention should be paid to the effect of a specific topic on the proportion of high and low MI-score collocations. As explained above, the MI-score favors less frequent and more specialised combinations; in particular, it privileges technical terms such as *carbon dioxide*. Thus the technical nature of a specific topic can influence the strength of collocations as measured by the MI-score. However, this factor is often hidden when only generalized patterns of, for example, MI-score rankings in whole corpora are considered. For example, consider the speaker from the Trinity Lancaster Corpus (Gablasova, Brezina, McEnery, & Boyd, 2015) who discussed the topic of social conditioning. She would be considered less nativelike in terms of collocational selection if we used the MI-score values derived from the BNC as our sole guide than another speaker who selected global warming as the topic because, unlike the topic of social conditioning, global warming is associated with a number of technical terms that by implication appear in the script such as *nuclear energy* (the BNC-derived  $MI = 7.37$ ), *toxic waste* ( $MI = 11.78$ ), *carbon dioxide* ( $MI = 15.05$ ), and *global warming* itself ( $MI = 14.1$ ). By contrast, *social conditioning* as a term has a relatively low MI-score of 5.29. We have to be mindful of factors such as the nature of the topic or a task (Ellis et al., 2015; Forsberg & Fant, 2010) when comparing and interpreting collocational patterns and strength to avoid somewhat arbitrary assessment of the collocations produced by L2 users (Wray, 2012). If these effects are not taken into consideration, we may be seeking psycholinguistic and developmental explanations for effects related to corpus composition and to a particular AM (e.g., MI-score). This is of particular relevance when using corpora that consist of language produced on a restricted

number of topics (e.g., ICLE has a set of recommended essay titles). Likewise, when evaluating variation between individual speakers, the effect of topic must be taken into account.

## Conclusion

This article focused on collocation research combining language learning and corpus linguistic perspectives, revisiting prior findings as well as suggesting a way forward. The main aim of this article was to make research on collocations more systematic, transparent, and replicable. To achieve this, we first need a good understanding of the mathematical and linguistic reasoning behind each AM in order to meaningfully apply the measures and explain the findings. Likewise, appropriate tools are needed (such as GraphColl, introduced in the article) that allow researchers to explore a full set of AMs in a replicable and transparent manner, as well as to create and apply their own measures. It was not possible to discuss all factors that play a role in identification of collocations or to provide a detailed description of all AMs; instead, the article demonstrated the principles that should underlie the choice, application, and interpretation of collocational measures in L1 as well as L2 production.

Final revised version accepted 19 November 2016

## Note

- 1 All of these examples occur in the BNC with the MI-score values above 20 and frequency above 10.

## References

- Ackermann, K., & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL)—A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12, 235–247. doi:10.1016/j.jeap.2013.08.002
- Arrpe, A., Gilquin, G., Glynn, D., Hilpert, M., & Zeschel, A. (2010). Cognitive corpus linguistics: Five points of debate on current theory and methodology. *Corpora*, 5, 1–27. doi:10.3366/cor.2010.0001
- Bartsch, S., & Evert, S. (2014). Towards a Firthian notion of collocation. *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information*. 2nd Work Report of the Academic Network Internet Lexicography, OPAL—Online publizierte Arbeiten zur Linguistik. Mannheim: Institut für Deutsche Sprache.
- Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, 26, 28–41. doi:10.1016/j.jslw.2014.09.004

- Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English*. New York: Longman.
- Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20, 139–173. doi:10.1075/ijcl.20.2.01bre
- Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1), 22–29.
- Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL-International Review of Applied Linguistics in Language Teaching*, 47, 157–177. doi:10.1515/iral.2009.007
- Durrant, P., & Siyanova-Chanturia, A. (2015). Learner corpora and psycholinguistics. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 57–77). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139649414.004
- Ebeling, S. O., & Hasselgård, H. (2015). Learner corpora and phraseology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 185–206). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139649414.010
- 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. doi:10.1017/S0272263102002024
- Ellis, N. C. (2014). Frequency-based accounts of second language acquisition. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 193–210). London: Routledge.
- 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, 375–396. doi:10.1002/j.1545-7249.2008.tb00137.x
- Ellis, N. C., Simpson-Vlach, R., Römer, U., Brook O'Donnell, M., & Wulff, S. (2015). Learner corpora and formulaic language in second language acquisition. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 357–378). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139649414.016
- Erman, B., Forsberg Lundell, F., & Lewis, M. (2016). Formulaic language in advanced second language acquisition and use. In K. Hyltenstam (Ed.), *Advanced proficiency and exceptional ability in second languages* (pp. 111–148). Boston: Walter de Gruyter.
- Evert, S. (2005). *The statistics of word co-occurrences: Word pairs and collocations*. (Doctoral dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.)
- Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 1212–1248). Berlin, Germany: Mouton de Gruyter. doi:10.1515/9783110213881.2.1212

- Forsberg, F., & Fant, L. (2010). Idiomatically speaking: Effects of task variation on formulaic language in highly proficient users of L2 French and Spanish. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 47–40). London: Continuum.
- Gablasova, D., Brezina, V., McEnery, T., & Boyd, E. (2015). Epistemic stance in spoken L2 English: The effect of task type and speaker style. *Applied Linguistics*. doi:10.1093/applin/amv055
- Gass, S. M., & Mackey, A. (2002). Frequency effects and second language acquisition. A complex picture? *Studies in Second Language Acquisition*, 24, 249–260. doi:10.1017/S0272263102002097
- Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5, 1–26. doi:10.1515/CLLT.2009.001
- González Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have?: The effects of frequency and amount of exposure. *International Journal of Applied Linguistics*, 166, 94–126. doi:10.1075/itl.166.1.03fer
- Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, 52, 229–252. doi:10.1515/iral-2014-0011
- Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 27–50). Amsterdam: John Benjamins. doi:10.1075/z.139
- Gries, S. Th. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 3–25). Amsterdam: John Benjamins. doi:10.1075/z.139
- Gries, S. Th. (2010). Dispersions and adjusted frequencies in corpora: Further explorations. *Language and Computers* 71, 197–212. doi:10.1075/ijcl.13.4.02gri
- Gries, S. Th. (2013). 50-something years of work on collocations: What is or should be next . . . . *International Journal of Corpus Linguistics*, 18, 137–166. doi:10.1075/ijcl.18.1.09gri
- Grosjean, F. (2010). *Bilingual: Life and reality*. Cambridge, MA: Harvard University Press.
- Handl, S. (2008). Essential collocations for learners of English: The role of collocational direction and weight. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 43–66). Amsterdam: John Benjamins. doi:10.1075/z.139
- Herbst, T. (2011). Choosing *sandy beaches*—collocations, probabemes and the idiom principle. In T. Herbst, S. Faulhaber, & P. Uhrig (Eds.), *The phraseological view of language. A tribute to John Sinclair* (pp. 27–57). Berlin, Germany: Walter de Gruyter.

- Howarth, P. (1998). Phraseology and second language proficiency. *Applied linguistics*, 19, 24–44. doi:10.1093/applin/19.1.24
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139524773
- McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge, UK: Cambridge University Press.
- Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins. doi:10.1075/scl.14
- Nguyen, T. M. H., & Webb, S. (2016). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research*, 1–23. Advance online publication. doi:10.1177/1362168816639619
- Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32, 130–149. doi:10.1017/S0267190512000098
- Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and Communication* (pp. 191–225). London: Routledge.
- Phillips, M. (1985). *Aspects of text structure: An investigation of the lexical organisation of text*. Amsterdam: North-Holland.
- R Core Team. (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Retrieved from: <http://www.R-project.org>
- Rebuschat, P., & Williams, J. (Eds.). (2012). *Statistical learning and language acquisition*. Berlin, Germany: Walter de Gruyter.
- Rychlý, P. (2008). A lexicographer-friendly association score. In P. Sojka & A. Horák (Eds.), *Proceedings of recent advances in Slavonic natural language processing, RASLAN* (pp. 6–9). Brno: Masaryk University.
- Schmitt, N. (2012). Formulaic language and collocation. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–10). New York: Blackwell. doi:10.1002/9781405198431.wbeal0433
- Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31, 487–512. doi:10.1093/applin/amp058
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.
- Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review*, 64, 429–458. doi:10.3138/cmlr.64.3.429
- Siyanova-Chanturia, A. (2015). Collocation in beginner learner writing: A longitudinal study. *System*, 53, 148–160. doi:10.1016/j.system.2015.07.003
- Siyanova-Chanturia, A., & Martinez, R. (2015). The idiom principle revisited. *Applied Linguistics*, 36, 549–569. doi:10.1093/applin/amt054
- Smadja, F., McKeown, K. R., & Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics*, 22, 1–38.



- Sosa, A. V., & MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word of. *Brain and Language*, 83, 227–236. doi:10.1016/S0093-934X(02)00032-9
- Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford, UK: Blackwell.
- Williams, G. (1998). Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, 3, 151–171.
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, UK: 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. doi:10.1017/S026719051200013X
- Wray, A. (2013). Formulaic language. *Language Teaching*, 46, 316–334. doi:10.1017/S0261444813000013
- Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics*, 32, 430–449. doi:10.1093/applin/amr011
- Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition*, 27, 567–595. doi:10.1017/S0272263105050254

## Supporting Information

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

**Appendix S1.** Data.

**Appendix S2.** Association Measures.

**Appendix S3.** Additional Results.