Taiwan Journal of TESOL
Vol. 18.1, 29-61, 2021 DOI: 10.30397/TJTESOL.202104_18(1).0002

ACQUISITION OF L2 COLLOCATION COMPETENCE: A CORPUS ANALYSIS OF EXCLUSIVITY, DIRECTIONALITY, DISPERSION AND NOVEL USAGE

Alvin Cheng-Hsien Chen

ABSTRACT

This study evaluates the development of L2 collocational competence in texts written by learners of differing proficiency levels, compared to native speaker collocation patterns from a reference corpus. We address: (1) whether learners develop their collocation competence as their proficiency grows; and (2) How is this development mediated by different aspects of collocability, i.e., exclusivity, directionality, and dispersion? Effective quantitative metrics based on the native corpus were assigned to each bigram type in L2 texts, covering important aspects of collocability. Correlations between the text-based average scores of each metric and L2 proficiency were analyzed to examine the development of collocability in each dimension. Our results show that exclusivity increases with learner proficiency. When directionality is considered, learners develop native-likeness in forward-directed word selection across all levels; backward competence, however, improves more markedly at advanced levels. Our analysis also suggests learners start to use less deviant collocation patterns but more domain-specific bundles as their proficiency grows.

Key Words: collocation, writing assessment, delta P, mutual information, inverse document frequency

INTRODUCTION

Phraseology has received widespread attention in language learning (Appel & Trofimovich, 2017; Gablasova, Brezina, & McEnery, 2017; Pawley & Syder, 1983; Wood, 2015; Wray, 2002). It can be broadly stated as a general linguistic phenomenon that words tend to co-occur as bundles of variable lengths, i.e., multiword sequences ranging from idiomatic expressions to semantically compositional sequences (Wood, 2015). Of particular interest to the present study are the two-word combinations, which have often fallen under the cover term of collocation. It is suggested that collocation competence is considered an essential component in native-like mastery of an L2. Collocation represents an initial, yet crucial stage, where learners start to develop their grammatical competence of concatenating lexical items into longer sequences for more sophisticated linguistic expression and social communication. Collocation itself, however, is an ambiguous term which has been operationalized by scholars from many different perspectives. A general definition of collocation may be traced back to a general linguistic observation, which says that some words tend to occur in the same neighborhood (Firth, 1957; Sinclair, 1991). The recurrence of pairs of words has therefore been a central criterion in defining collocations.

While recurrence may seem an intuitive criterion for defining collocation, scholars differ in their approach to deriving a more restricted set of qualifying features for collocations. For example, collocation is sometimes used more restrictedly to refer to word combinations which have little semantic transparency (Nation, 2001; Nesselhauf, 2005), such as idioms (e.g., *spill the beans*) or fixed expressions (e.g., *nuts and bolts*). Alternatively, collocations can also refer to word combinations of relative semantic transparency, such as *strong coffee, heavy smoker*. They are uniquely defined as collocations due to the fact that one of the words in the combinations is highly constrained to this bundle with its unique semantics (e.g., *strong* and *heavy*). Collocations can also be defined even more broadly as word combinations that habitually co-occur, whose semantics can be fairly transparent (Biber & Conrad, 1999; Laufer & Waldman, 2011; Simpson-Vlach & Ellis, 2010; Sinclair, 1991): for example, *strong man* or *heavy load*.

This study adopts this broader co-occurrence based approach to collocation, and regards recurring word combinations as collocations. Most importantly, we subscribe to a graded view of collocation, by treating word combinations as bundles of varying conventionality depending on the degree of recurrence. In other words, collocation is considered not a *categorical* feature but a *quantitative* feature of a two-word sequence, which is defined based on the sequence's corpus-based distributional properties. On this continuum one may see word combinations whose meaning is semantically compositional based on their parts at one end, as well as idioms or fixed expressions whose meaning is fully opaque at the other extreme. Syntactically, collocations can be grammatically legitimate phrases, fully predictable from phrase-structure rules, or structurally fragmented. These two-word sequences can be more or less collocation-like depending on how "frequently" they occur in the corpus. Therefore, in this study, we use "collocation" in its most general sense to cover any type of habitually occurring word combination, to which may be ascribed a range of different terms depending on the research paradigm adopted, including lexical bundles, multiword expressions, CollGrams, and ngrams. Wray (2002, p. 9) has identified more than fifty different terms used in the previous literature for this general concept of formulaic language. In the following sections, we discuss different methods of defining the construct of "recurrence" (i.e., how frequently the sequences occur), which is greatly connected to the "formulaicity" of multiword combinations.

Operationalizing Collocations

A classic way of identifying collocations is to rely on proficient native speaker intuition (Laufer & Waldman, 2011; Nesselhauf, 2005). A collocation dictionary may be consulted for defining correct collocations. Collocations according to this approach have two important characteristics. First, collocation is considered a categorical property of a multiword unit (i.e., a two-word sequence is either collocation or non-collocation). Second, intuition-based collocation lists often include word combinations that are not semantically compositional (i.e., idiomatic expressions). In this dictionary-based approach, the assessment of learners' collocation knowledge often relies on a quantitative study of the frequencies and functions of these dictionary-derived or intuition-based collocations in L2 productions. More uses of these dictionary-listed collocations are indicators of L2 collocation competence. Word combinations in L2 productions that depart from these intuition-based collocations may be considered incorrect (or "deviant", to use Laufer & Waldman's [2011, p. 654] term).

Altenberg and Granger (2001) analyzed advanced learners' use of the high-frequency verb *make* by comparing its frequencies in a learner corpus and a native writer corpus. They categorized the collocations of *make* into eight functional types based on native judgement. In their observations, learners consistently show learning difficulty with delexical (e.g., *make a decision, make a reform*) and causative uses (e.g., *make someone believe something, make something possible*) of *make* (i.e., learners underuse *make* in these functions). Examining the *verb + noun* combinations of high-frequency nouns, Laufer and Waldman (2011) assessed the correctness of these combinations in the learner texts by consulting the collocation dictionaries. Their findings suggest that advanced learners did not show more uses of correct collocations.

Easy access to large corpora data has driven a distribution-based, bottom-up approach to research on collocations and phraseology (Ädel & Erman, 2012; Bestgen, 2017; Bestgen & Granger, 2014; Crossley & Salsbury, 2011; Durrant & Schmitt, 2009; Ellis, Simpson-Vlach, & Maynard, 2008; Leńko-Szymańska, 2014). Proficient native speaker intuition can be more reliably estimated using a large representative native corpus. Whether a two-word unit is a true collocation or not may now be an empirical question, which can be quantitatively answered given its distribution in the reference corpus. The semantic compositionality of the multiword unit may be less crucial in this distribution-based research paradigm. One of the most comprehensive distributional features is the co-occurrence frequency of two-word bundles in the corpus. These frequency-based collocations include phrases that are both semantically opaque and semantically compositional. This frequency-based approach to collocations provides two possible ways to analyze L2 collocational competence.

On the one hand, the term, collocation, is sometimes used more broadly to cover multiword units beyond two-word sequences that satisfy strict distributional criteria, such as frequency and range. Sequences meeting these criteria are often referred to as lexical bundles (Biber & Conrad, 1999; Biber, Conrad, & Cortes, 2004). Research on lexical bundles often focuses on sequences of more than two words, typically four-word bundles (Ädel & Erman, 2012; Biber et al., 2004; Cortes, 2004; Hyland, 2008). This approach often takes a cross-sectional design to analyze phraseological development by comparing the differences of the bundles observed in two contrastive corpora, varying in at least one external criterion, such as L2 proficiency (Appel & Wood, 2016), discipline (Hyland, 2008), register (Biber et al., 2004), or publishing experience (Chen & Baker, 2010). The structural or functional differences in the use of bundles are therefore connected to the distinctive feature of the two corpora. This study will limit our discussion of collocation competence to two-word sequences only.

On the other hand, the distribution-based approach provides the possibility of analyzing collocations as a graded property of the multiword units: the higher the co-occurrence frequency is, the more formulaic it is.

Every two-word bundle can be assessed in terms of its collocability based on its distribution in a representative native corpus. These distributional metrics in turn can be utilized to evaluate the bundles used by learners. That is, these distributional metrics informed by native corpora can serve as an effective measure to assess the native-likeness of each multiword sequence used by learners (Bestgen, 2017; Bestgen & Granger, 2014; Crossley & Salsbury, 2011; Durrant & Schmitt, 2009), and average scores of these bundles' distributional metrics can be generated to capture the degree of formulaicity in either each L2 text, or of the whole collection.

Corpus-Based Distributional Metrics

Two types of corpus-based distributional statistics have been commonly used in operationalizing the formulaicity, or "native-likeness", of collocations: frequency, and statistical associations. Frequency concerns the most intuitive distributional evidence that words co-occur frequently, and this can be taken as an indicator of formulaicity. For instance, Crossley and Salsbury (2011) analyzed the use of two-word bundles in the spoken English production of learners of different L1 backgrounds across a whole year. Using the Santa Barbara Corpus of Spoken American English as a reference corpus, Crossley and Salsbury first identified all bigrams used in L2 spoken productions that were also present in the reference corpus and compared the correlations of the normalized frequencies of these bigrams in both corpora. They assumed that the more the frequency distribution of these shared bigrams in the learner production approximated the native speaker use, the more accurate the bigram use was. According to their findings, the correlation increases significantly with the time spent in English learning. Kyle and Crossley (2015) also estimated formulaicity of all bigrams and trigrams in L2 texts using the frequency-based scores of these bundles in the British National Corpus (BNC). They found that the ngram frequency-based indices show strong positive correlations with a learner's speaking and lexical proficiency scores, accounting for 22.3-35.2% of the variance.

While frequency of a multiword unit can be an effective and useful metric for the formulaicity of the bundle, statistical associations have been used more often in research on L2 collocation acquisition because the significance of the frequency of a multiword unit may need to be evaluated in terms of the frequency of its parts (Evert, 2009). Two widely-used association measures are mutual information (MI) and *t*-scores

(Gablasova et al., 2017; Hunston, 2002). Focusing on two-word sequences co-occurring in a *modifier* + *noun* construction, Durrant and Schmitt (2009) analyzed the lexical associations of these bigrams using MI scores and *t*-scores, which were computed based on the British National Corpus (BNC). In their study, L2 learners tended to underuse bigrams of high mutual information scores, which were often low in frequency. Bestgen and Granger (2014) further extended the analysis of bigrams to all contiguous two-word sequences in L2 texts. Every L2 text was assigned three indices: the mean MI score and mean *t*-score of all bigrams in the text and the Pabsent rate, i.e., the proportion of unseen bigrams in the reference corpus. Their analysis showed that the MI mean scores of L2 texts positively correlated with the human ratings of writing assessment, but higher Pabsent rates were connected to lower ratings. Also, their findings suggest that the formulaicity average scores (i.e., MI) based on bigram *types* correlate with the human text ratings in a more significant way than those based on bigram *tokens*. Their findings support the hypothesis that more proficient learners do produce bigrams that are more “formulaic” (as defined by higher average MI scores). The metric of *t*-score did not seem effective in predicting the human text quality ratings; this may be partly due to its confounding strong association with high frequency words. Bestgen (2017) further suggests that the phraseological metrics provided by CollGrams out-performed single-word lexical measures of diversity and sophistication in predicting the human text quality rating. Similarly, Kyle, Crossley, and Berger (2018) observed that corpus-based indices related to phraseology association strength and frequency are able to account for almost 28% of the variance of human ratings on learner samples.

This Study

Studies adopting a categorical view of collocation tend to analyze the learner’s overuse, underuse, correctness, and functional salience of the collocations (Laufer & Waldman, 2011; Nesselhauf, 2005) while those adopting a graded view of collocation may evaluate the development of “formulaicity” in learners’ use of multiword items as a numeric trend (Bestgen, 2017; Bestgen & Granger, 2014; Kyle et al., 2018). While both approaches have provided many insights into L2 collocational knowledge, there are two important gaps in the literature on the L2 collocation competence in terms of their use of two-word combinations. First of all, studies adopting a categorical approach often focus on structurally-dependent collocations. Despite different ways of defining collocations, most of the previous studies often adopt a particular constructional schema as a basis for the analysis of collocations, e.g., *modifier* + *noun* in Durrant and Schmitt (2009), *noun* + *adjective* (of L2 Italian) in Siyanova-Chanturia (2015), or *verb* + *noun* in Laufer and Waldman (2011).

One major disadvantage of analyzing structurally dependent collocations is that the conclusions may not necessarily be generalizable to the learner's overall collocation competence. Sometimes contradictory results may be obtained in different studies. For example, Laufer and Waldman (2011) analyzed the *verb* + *noun* combinations used by learners and found that learners did not show a noticeable increase in use of collocations as their proficiency grew. Siyanova-Chanturia (2015) conducted a longitudinal study on the *noun* + *adjective* combinations used by beginner learners of Italian throughout an intensive course. They first computed the MI scores of the bigram pairs observed in a learner corpus, comparing with a native Italian reference corpus. They observed that at the end of the course, learners produced more *noun* + *adjective* combinations with higher frequencies and MI scores, suggesting a development in collocational knowledge in beginner learners of L2 Italian. Both Laufer and Waldman (2010) and Siyanova-Chanturia (2015) examined collocations that were structurally dependent in different types of construction. Their contradictory findings may be partly attributed to the fact that collocational development based on particular syntactic structures may not necessarily generalize to the overall development of collocation competence. Moreover, their different operational definitions of structurally-dependent collocations may further render their findings less comparable. Siyanova-Chanturia (2015) used a distribution-based bottom-up method while Laufer and Waldman (2011) adopted a more top-down dictionary-based approach. Given their different methodological emphases, it remains unclear to readers of the two papers whether learners show a clear development in their collocational knowledge as their proficiency grows in terms of all types of two-word combinations.

The second important gap is that previous corpus-based research on L2 collocation often neglects several important dimensions of collocability in phraseological development. While frequency information has been one of the most intuitive distributional properties provided by corpus data, it can be misleading and may therefore need to be assessed by considering other important aspects of the distributional properties of the linguistic units. In particular, Gablasova et al. (2017) point out that the distributional properties of linguistic units may need to consider three important dimensions of collocability: exclusivity, dispersion, and directionality. Exclusivity concerns the statistical significance of the extent to which the words' co-occurrence is beyond the expected frequency. Dispersion is the evenness of distribution of the multiword unit in a corpus. Directionality highlights the fact that words in a bundle are not always attracted to each other with equal strength. When learners develop their collocation competence, they may develop their sensitivity to this multifaceted nature of the distributional properties (Ellis, O'Donnell, & Römer, 2014; Ellis & Ogden, 2017). Corpus data can provide relevant distributional metrics for us to further examine the distributional differences of the collocations.

In this study, we use the Corpus of Contemporary American English (COCA) as our source of distributional metrics. Take the following two bigrams, *Monday night* and *excellent swimmer,* for example. In COCA, there are 2314 tokens of *Monday night* and 12 tokens of *excellent swimmer*. The raw frequencies of these two bigrams may give the impression that *Monday night* is more formulaic than the other. However, based on several corpus-based quantitative metrics to be further introduced in Method, these two bigrams can be compared more comprehensively by considering the exclusivity, directionality, and dispersion of their distributional properties.

To begin with, when considering the bigram's lexical associations, we can analyze the property of exclusivity of these bigrams in addition to their frequencies. Based on the mutual information scores of *Monday night* (MI = 7.83) and *excellent swimmer* (MI = 7.70), the lexical items of these two bigrams are almost equally exclusive to each other even though their frequencies differ by two orders of magnitude. In other words, these two bigrams may be equally important as conventional expressions in English in terms of the exclusivity aspect of the bigram distribution.

Second, when adopting lexical associations with directionality (See delta P in Collocability Metrics), we can analyze whether the lexical items in these two bigrams are attracted to each other in a symmetrical way. According to the delta P scores (See Collocability Metrics for a step-by-step computation) of these two bigrams, *Monday night* is a forward-directed collocation, where the first word, *Monday*, more strongly prompts the second word, *night*; in contrast, *excellent swimmer* is a backward-directed collocation, where the second word, *swimmer*, more strongly prompts the first word, *excellent*. Therefore, these two bigrams may differ in the relative strengths of their forward-directed and backward-directed lexical associations.

Studies we have reviewed so far (Bestgen, 2017; Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Siyanova-Chanturia, 2015) seem to have stressed mainly the first dimension, exclusivity, when analyzing the development of L2 collocation competence. More specifically, they have mostly adopted non-directional association measures, such as MI or *t*-scores. These association measures do not address to what extent the development of the L2 collocation knowledge may be mediated by the directionality of collocability. It is therefore unclear whether learners develop collocation competence differently in terms of their native-likeness in forward and backward word selection. Learners may develop collocation competence by using word combinations that are more native-like in terms of forward-directed temporal relations between words. For example, when using the word *apply,* learners may demonstrate a forward-directed collocation knowledge if they choose a preposition *for* after *apply*. On the other hand, learners may develop their collocation knowledge by using word combinations that are more native-like in terms of backward-directed temporal relations of words. For example, given a word *home*, learners may demonstrate the collocation knowledge when choosing the preposition *at* before *home*.

Finally, *Monday night* and *excellent swimmer* may also differ in their dispersion. According to their distribution in COCA, *Monday night* is a bigram which is more widely-dispersed in different documents than *excellent swimmer*: the former is found in 119 different documents in the entire corpus while the latter is found in only 11 documents. Lexical association measures (i.e., MI, *t*-score, delta P) would not inform the degree of dispersion of the collocation, which may however play a role in the development of L2 collocation competence. While previous studies have identified a positive relationship between L2 proficiency and the average MI scores of the two-word sequences used by the learners (Bestgen, 2017; Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Siyanova-Chanturia, 2015), it remains unclear how this increase of exclusivity in two-word sequences may be mediated by their dispersion rates. We may wonder whether learners also develop collocation competence by acquiring word sequences that are more domain-general (i.e., sequences that are widely-dispersed in different documents) at the beginning and mastering ones that are more domain-specific (i.e., sequences that are more centralized in particular sets of documents) in later learning phases.

The objective of this study is thus to bridge these gaps by (1) assessing the collocational development not in a particular morphosyntactic setting but in all the two-word sequences used by learners, and (2) examining whether learners develop their collocation competence as their proficiency level grows in terms of the exclusivity, directionality and dispersion of collocability. To address these important issues, we follow the distribution-based approach to collocation and take a large representative native corpus as a "proxy for native speaker intuition" (Bestgen, 2017, p. 66), from which a range of quantitative metrics will be utilized to assess L2 collocation knowledge. We will utilize not only commonly used association measures, such as MI and *t*-scores, to assess the development in exclusivity, but also adopt an effective directional association measure (Ellis, 2006; Gries, 2013), delta P (DP), to see if learners develop their collocation knowledge in different directions. Also, we will use a useful metric of dispersion, inverse document frequency, to address the issue of dispersion. These effective distributional metrics will be computed based on the Corpus of Contemporary American English (Davies, 2012), a proxy for the proficient speaker's intuition in co-selection of words in varying scenarios. By considering different dimensions of collocability, we hope to provide a more comprehensive picture of L2 development in collocation competence.

METHOD

Data

This study analyzed the L2 texts collected in the International Corpus Network of Asian Learners of English V2.0 (ICNALE) (Ishikawa, 2013). This learner corpus includes around 2 million words from essays and monologues produced by both L1 writers and L2 English learners from different countries of Asia. We analyzed all essays written by L2 learners, which amounted to 5200 essays. For each L2 text, ICNALE annotated the proficiency level of the learner using the reference points of the Common European Framework of Reference for Languages (CEFR). The learner proficiency levels were defined based on external criteria, using standardized English proficiency tests (TOEIC, TOEFL, or IELTS) or an objective vocabulary size test (Nation & Beglar, 2007). In ICNALE, the original CEFR B2, C1, and C2 were collapsed into B2+ and the original B1 was subdivided into B1_1 and B1_2 in order to better represent the largest group of Asian intermediate-level learners (cf. Ishikawa, 2013). Thus, learners were grouped into four proficiency levels: A2, B1_1, B1_2, and B2+. Table 1 shows the distribution of texts in all levels.

Table 1

Data Distribution of ICNALE 2.0

Proficiency Level	**Number of Texts**	**Number of Words**	**Mean Text Length**	**SD**
A2	960	216479	225.5	22.95
B1_1	1904	437904	229.99	25.92
B1_2	1872	439631	234.85	28.49
B2+	464	111916	241.2	29.94

The present study used the Corpus of Contemporary American English (COCA) (Davies, 2012) as the reference native corpus for the estimation of a range of collocability features. COCA comprises 560-million-word texts of American English, which are equally divided among spoken, fiction, popular magazine, newspaper and academic genres from 1990 to 2012. Given its size and representativeness, this corpus can serve as a yardstick by which an ideal proficient native speaker intuition of collocation knowledge can be quantitatively estimated. All the collocability metrics in this study were based on COCA. All data preprocessing and statistical computation was done with self-developed scripts written in R.

Data Preprocessing

To generate proper estimates of collocability metrics, the reference corpus was preprocessed as follows. All HTML/XML tags in the corpus were removed. Raw texts in each corpus file were segmented into chunks by taking as the delimiters all non-word tokens consisting of symbols except for word-internal characters (i.e., the hyphen - and the apostrophe '). This was to ensure that the later extraction of contiguous two-word sequences did not span the boundaries of sentences and punctuation marks. All contiguous two-word combinations were extracted from each corpus file. Bigrams containing numbers were removed. All bigrams extracted were normalized into lower-case letters. To control for the minimum frequencies and the dispersion of the bigrams included in the reference corpus model, only bigrams of raw frequency > 10, occurring in at least five different documents in the entire COCA were included.

After data preprocessing, we identified 2,334,463 bigram types from COCA. For each type, we further computed several distributional metrics that characterized the three aspects of their collocability (i.e., exclusivity, directionality, and dispersion). The next section will introduce the statistical metrics and their computation for each dimension of the collocation competence.

Collocability Metrics

As collocation has been operationalized in many different ways, this study adopts a corpus-based method, and relies on distributional statistics of words, which are less subjective compared to methods based on native intuition judgements. After data preprocessing, the reference corpus provided the necessary distributional statistics for estimating different aspects of collocability for every bigram type in COCA. The frequency information is arranged in a contingency table, as Table 2. This study investigated L2 collocation competence from four important perspectives, each of which was quantitatively measured utilizing the distributional statistics of Table 2 informed by the native corpus. The following sections present the mathematical computations of each metric.

Table 2.

Contingency Table for W_1W_2 Collocability Metrics Computation

	W_2	$\neg W_2$	
W_1	O_{11}	O_{12}	R_1
$\neg W_1$	O_{21}	O_{22}	R_2
	C_1	C_2	N

Notes. *O* refers to the observed frequencies of each cell; *R* refers to the sums of the rows; *C* refers to the sums of the columns. O_{11} refers to the co-occurrence frequency of the two words; O_{12} refers to the frequency of W_1 in the absence of W_2; O_{21} refers to the frequency of W_2 in the absence of W_1; O_{22} refers to the frequency of all the other two-word sequences that are not W_1W_2.

Exclusivity

First, we analyzed the exclusivity of the two-word sequences using MI and *t*-scores. Given a potential bigram, W_1W_2, observed in COCA, we estimated its exclusivity using the frequency distributions of its sub-units. The association measures for a bigram W_1W_2 were computed using (1) and (2), which are based on Evert (2008).

(1) $\mathrm{MI}(W_1,W_2) = \log_2 \frac{p(W_1,W_2)}{p(W_1)\times p(W_2)}$

(2) $t\text{-score}(W_1,W_2) = \frac{p(W_1,W_2) - p(W_1)\times p(W_2)}{\sqrt{p(W_1,W_2)}}$

In the formulas, the $P(W_1,W_2)$ refers to the joint probability of W_1 and W_2 in COCA; $P(W_1)$ and $P(W_2)$ refer to the respective probabilities of W_1 and W_2 in COCA.

Directionality

The rationale behind directionality is that words in a collocation may not be attracted to each other in a symmetrical way. Delta P (DP) is an effective metric for capturing a directional association between a cue and an outcome (Ellis, 2006; Gablasova et al., 2017; Gries, 2013). It is a normalized conditional probability of an outcome given a cue, i.e., *P(outcome|cue)*, which considers the potential impact of the conditional probability of an outcome in the absence of the cue, i.e., *P(outcome |¬cue)*.

This metric can be used to produce directional lexical associations of any two-word sequence, i.e. W_1W_2. When W_2 is taken as the outcome and W_1 as the cue, a forward-directed DP can be computed using the formula in (3); on the other hand, when W_1 is taken as the outcome and W_2 as the cue, a backward-directed DP can be computed using the formula in (4).

(3) Forward Delta P of W_1W_2:

$$Delta\ P = P(W_2|W_1) - P(W_2|\neg W_1) = \frac{O_{11}}{R_1} - \frac{O_{21}}{R_2}$$

(4) Backward Delta P of W_1W_2:

$$Delta\ P = P(W_1|W_2) - P(W_1|\neg W_2) = \frac{O_{11}}{C_1} - \frac{O_{12}}{C_2}$$

We generated directional DPs, forward and backward, for all bigrams in COCA, amounting to 2,334,463 different bigram types. These adjusted conditional probabilities can be useful indicators of the native-like intuition in forward- or backward-directed word co-selection. For example, according to the forward DP based on COCA, the top five words that most likely follow the first-person pronoun *I…* are *am*, *think*, *do*, *was*, and *have*. If the cue is different, e.g., *you…,* then the native-like intuition for forward word selection may predict a different set, i.e., *know, are, can, have,* and *do*. Similarly, a native-like intuition for backward word selection would predict that the top five words that most likely come before *home* are *at, go, back, his,* and *come*. A different cue word like *house* would lead to a different set of words likely preceding the cue, i.e., *the, white, 's, a,* and *my*. It is hypothesized that more advanced learners may perform the co-selection of words more similarly to native-speaker intuition. This study makes a step further examining whether directionality in word co-selection plays a role.

Dispersion

Dispersion is an effective notion in assessing the learner's use of collocations in terms of the domain-specificity of the collocations. It is posited that learners may start to acquire collocations that are common in general situations (i.e., those that are high in dispersion) and start to acquire those that are used in particular domains (i.e., those that are low in dispersion) in later learning phases. For every bigram type in COCA, we computed a useful metric, inverse document frequency (IDF), which was inspired by its effectiveness in information retrieval (Manning & Schütze, 1999). A comprehensive review of various dispersion metrics can be found in Gries (2010). IDF is calculated as follows.

(5) Inverse Document Frequency (IDF) = $log\left(\frac{d}{N}\right)$

In (5), the *d* refers to the number of documents in COCA where the bigram is observed; *N* refers to the total number of the documents in COCA. If a bigram occurs in every document of COCA, the IDF would be 0. If a bigram is concentrated in only a few documents in COCA, its IDF would increase.

Unseen Rates

All the aforementioned metrics were targeted toward bigrams used by learners that were also present in the native reference corpus. That is, the metrics analyzed bigrams that were found in both L2 texts and COCA. Following CollGrams (Bestgen, 2017; Bestgen & Granger, 2014), we considered as well the rates of bigrams that are absent in the reference corpus in the learner's production. An unseen bigram may be significant in two important senses. On the one hand, an unseen word combination may be an ungrammatical or incongruent sequence in English (i.e., a deviant word combination); on the other hand, a novel combination may suggest a learner has mastered creative use of collocation to some extent.

Research Questions

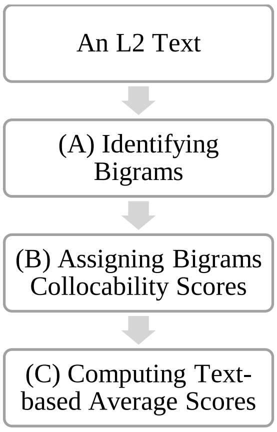
An L2 Text
(A) Identifying Bigrams
(B) Assigning Bigrams Collocability Scores
(C) Computing Text-based Average Scores

Figure 1. Flowchart of the data analysis

The collocability metrics of bigrams collected from COCA were used as a reference list to analyze the acquisition of L2 collocation knowledge in ICNALE texts. Figure 1 provides a flowchart of our data analysis. First, each L2 text in ICNALE was preprocessed with the same procedure as for the COCA to create a list of bigram types used by each learner (Step [A] in Figure 1). Table 3 shows the number of bigram types for the L2 text collections of each proficiency level. Among the four levels, B2 learners show the most use of the bigrams that were present on the COCA reference list, with on average 62.13 bigram types per text.

Table 3

Number of Bigram Types Observed in L2 Texts by Proficiency Levels

Level	Bigram Types	Number of Texts	Number of Bigram Types Per Text
A2	39459	960	41.10
B1_1	62045	1904	32.59
B1_2	70766	1872	37.80
B2+	28830	464	62.13

Notes. Bigram types refer to the L2 bigrams that are present in the reference list identified in the native corpus, i.e., Corpus of Contemporary American English.

After identifying the bigram types of each L2 text, we assigned each L2 bigram type five collocability scores. As introduced in Method, these scores were computed based on the distributional properties of these bigrams in COCA, highlighting different aspects of collocability—MI scores and *t*-scores for exclusivity, forward and backward DP for directionality, IDF for dispersion (Step [B] in Figure 1). Finally, we computed the text-based mean scores of each collocability metric for each L2 text by calculating the average scores of all the L2 bigram types (Step [C] in Figure 1). The proportion of bigram types that were absent in the reference native corpus was also computed for each L2 text. Therefore, each L2 text had six collocability metrics in total.

The main objective of this study was to examine whether the text-based mean collocability scores increase with learner proficiency. Two questions were addressed in this study:

- Do learners develop their collocational knowledge in two-word sequences as their proficiency grows?

- How is the development of collocation competence mediated by different aspects of collocability, i.e., exclusivity, directionality, dispersion, and novelty (use of unseen collocations)?

Learner proficiency was defined as an ordinal dependent variable LEVEL with four values: A2, B1_1, B1_2, and B2+. We analyzed how LEVEL correlates with collocability metrics on different dimensions, including the exclusivity (measured by MI and *t*-score), directionality (measured by forward and backward DP), dispersion (measured by IDF) and novelty (measured by unseen bigram rate). Depending on the matching degrees of the statistical assumptions of each metric, appropriate statistical methods were used to determine the significance of the phraseological development.

ANALYSES AND RESULTS

Exclusivity

Exclusivity was operationalized using non-directional association measures, MI and *t*-score. As neither metric satisfied the statistical assumptions of normality and variance homogeneity, we adopted two non-parametric Kruskal-Wallis tests on the variation of MI and *t*-score in relation to LEVEL. Our results show that both metrics were significantly affected by LEVEL (*t*-score: $H(3) = 28.24$, $p < 0.01$, $r = -0.06$; MI: $H(3) = 275.95$, $p < 0.01$, $r = -0.23$). We conducted a post-hoc Jonckheere-Terpstra test, which is a useful trend analysis for the ordered pattern to the medians of the groups compared (cf. Ch 15 in Field, Miles, & Field, 2012). The Jonckheere-Terpstra tests revealed a significant trend in MI scores: as learners progress to more advanced levels, the MI scores increase ($J = 5624700$, $p < 0.01$); however, no significant trend was found in *t*-scores. As shown in Figure 2, learners show a clear growing trend in the exclusivity of collocability measured by MI; the tendency measured by *t*-score may be less conclusive. We will come back to this point in the next section.

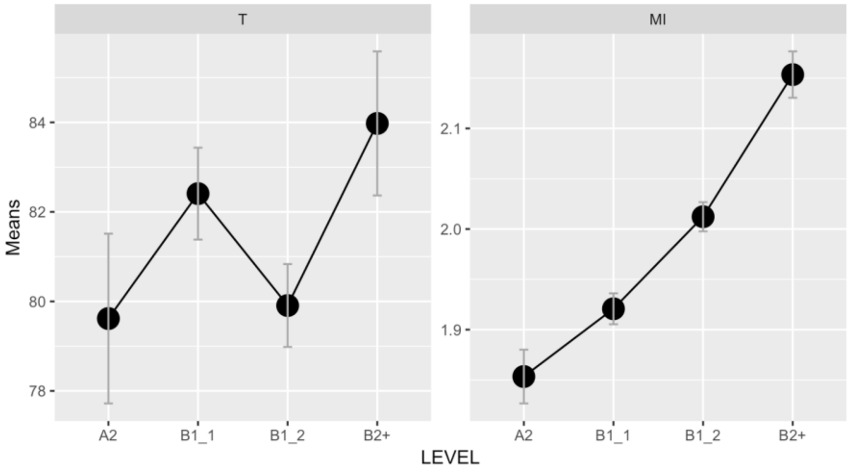

Figure 2. Means and 95% confidence intervals of MI and *t*-scores by LEVEL

Directionality

According to a Kolmogorov-Smirnov test, DP values of the L2 texts did not deviate from the normal distribution significantly ($D = 0.0159$, $p > 0.01$). In the analysis of DPs, each bigram type used in L2 texts was assigned two directional metrics, forward and backward DP. Mixed design ANOVA was used to analyze the variation of DP values in relation to its directionality (DIRECTION) and the L2 proficiency (LEVEL), with the former as a within-subject, the latter as a between-subjects factor. The model also included the interaction between LEVEL and DIRECTION. Polynomial orthogonal contrasts were used for *post-hoc* analyses. Results are shown in Table 4.

Table 4

ANOVA Table for Directionality

Model	**df**	**AIC**	**BIC**	**logLik**	**L.Ratio**
Intercept	4	-68464.23	-68435.23	34236.11	NA
LEVEL	7	-68547.87	-68497.12	34280.93	89.64*
DIRECTION	8	-69593.66	-69535.66	34804.83	1047.79*
LEVEL × DIRECTION	11	-69754.34	-69674.59	34888.17	166.68*

Notes. * = $p < 0.001$

The main effect of LEVEL suggests that DP varies significantly across different proficiency levels. Figure 3 plots the DP mean scores of learners of each proficiency level, showing a general increasing trend in DP with learner proficiency. The general tendency of DIRECTION is that learners use collocations of higher backward DP values on average. Most importantly, there was a significant interaction between LEVEL and DIRECTION. The *post-hoc* analysis suggests that the linear trends across different proficiency levels are significantly different for forward and backward DPs (β = -0.004, *SE* = 0.00004, $t(5196) = -10.2571$, $p < 0.01$, $r = 0.14$). We computed the effect size (r) of the interaction based on the focused contrast using the formula below (Field et al., 2012, p. 640):

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

Figure 5 provides a graphic illustration of the interactional effect. While the backward DP is higher than the forward DP on average, learners seem to demonstrate a more stable growing pattern in the forward DPs. The development of the backward DP may be less prominent until learners reach a more advanced level (e.g. from B1_2 to B2+).

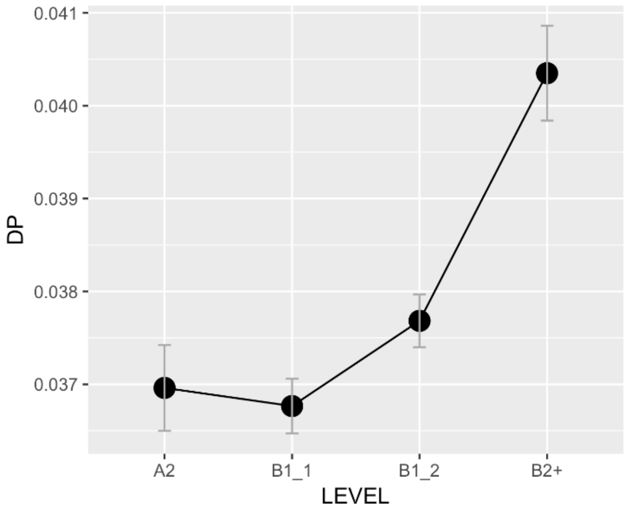

Figure 3. Means and 95% confidence intervals of DP by LEVEL

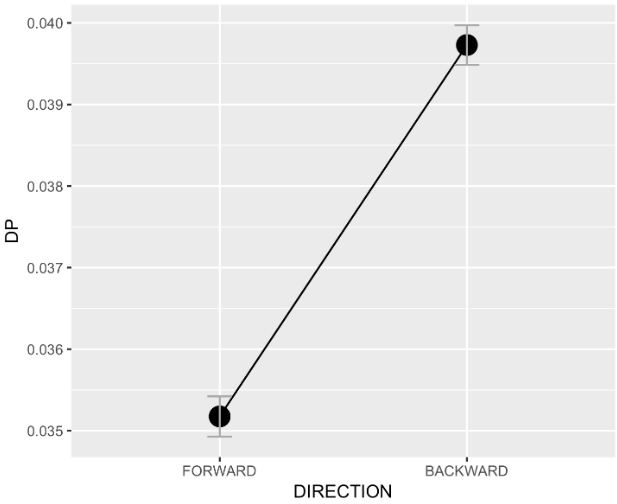

Figure 4. Means and 95% confidence intervals of DP by DIRECTION

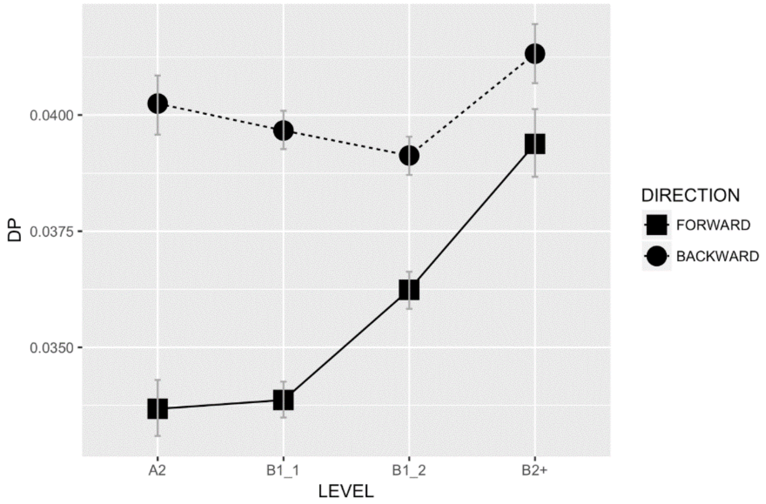

Figure 5. Interaction plot of LEVEL X DIRECTION on DP

Dispersion

Dispersion of collocations was evaluated using the IDF. A higher IDF mean score for an L2 text may suggest that the two-word sequences used by the learner are on average more concentrated in particular sets of documents in the corpus, i.e., more domain-specific (or idiosyncratic). It is suggested that the acquisition of domain-specific collocations may emerge more markedly in more proficient learners.

The IDF values in our data met the statistical assumption of normality but violated the assumption of variance homogeneity. As ANOVA is generally robust to this variance violation when the sample size is large, it was used to analyze the differences of IDF among the four proficiency levels. Our results show that LEVEL has a significant effect on IDF with a small effect size ($F(3, 5196) = 24.39$, $p < 0.01$, $\omega^2 = 0.01$). The *post-hoc* comparisons suggest only a significant growth in IDF when learners develop from B1_1 to B1_2, as illustrated in Figure 6.

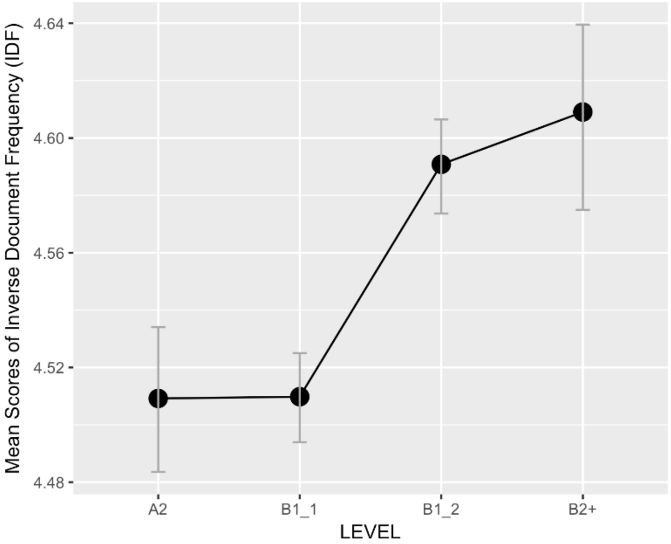

Figure 6. Means and 95% confidence intervals of Inverse Document Frequency (IDF) by LEVEL

Unseen Rates

Unseen rates (URs) were a simple percentage, showing the proportion of the two-word sequences in L2 essays that were absent from the native corpora, i.e. COCA. Because a large representative native corpus may be expected to have included most salient collocation possibilities in English, an unseen bigram may be either an ungrammatical word combination or a highly creative use. As the distribution of URs in our data violated the statistical assumptions of normality and variance homogeneity, we adopted a non-parametric Kruskal-Wallis test to analyze the LEVEL effect on the UR variation.

The result shows that URs were significantly affected by LEVEL with a small effect size ($H(3) = 14.82$, $p < 0.01$, $r = -0.04$). The Jonckheere-Terstra test revealed a significant linear negative relationship between UR and LEVEL ($J = 4504600$, $p < 0.01$), indicating that learners show smaller UR on average as their proficiency grows. The trend of UR variation by LEVEL was given in Figure 7.

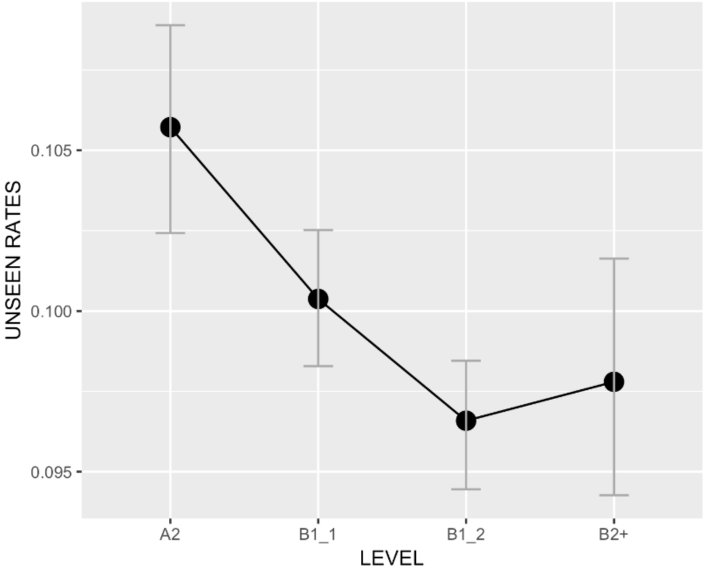

Figure 7. Means and 95% confidence intervals of Unseen Rates (UR) by LEVEL

DISCUSSION

This study examined the development of learner collocation competence by analyzing four aspects of collocability: directionality, exclusivity, dispersion, and novelty. Our analysis on exclusivity aligns with previous studies, showing a positive relationship between the exclusivity of the two-word sequences used in L2 texts and learner proficiency levels. Also, it is suggested that MI may be a more effective metric in showing the growth in exclusivity; *t*-scores did not reveal a clear linear growth across different proficiency levels, and this has also been observed in CollGrams-based studies (Bestgen & Granger, 2014).

When a metric does not positively correlate with the proficiency level, two interpretations are possible: (1) learners do not develop the construct measured by the metric as their proficiency grows, or (2) the metric is not an effective operational measure for capturing the development of the construct. In this study, we are more inclined to adopt the latter interpretation for the *t*-scores. As most of our metrics in the other collocability dimensions all point to a positive growth in learners' collocation competence, we suggest that exclusivity based on *t*-scores may be confounded by the fact that *t*-scores are often sensitive to high-frequency words (Evert, 2009; Gablasova et al., 2017; Hunston, 2002). Bestgen and Granger (2014) evaluated the bigrams used by learners with the same sets of association measures, MI and *t*-score, computed based on COCA. They analyzed the correlation between the average MI and *t*-score of all bigrams and the human ratings of each learner text. Different from our study, their text-based mean scores were analyzed both in terms of bigram tokens and types. Their results clearly suggest that the correlation between *t*-scores and the text ratings was substantially higher (from r_{type} = 0.03 to r_{token} = 0.11) when it was computed based on tokens than when based on types (Bestgen & Granger, 2014, p. 37). We posit that *t*-scores may not be an effective phraseological metric in the assessment of collocation competence. What MI could give us is a more conclusive pattern: learners tend to use bigrams that are more strongly associated as their proficiency grows.

Furthermore, our analysis has also identified different patterns of development in collocation competence in terms of the directionality of collocability. Learners show a steady linear growth in the forward collocability across different proficiency levels, but this tendency is obscured in backward collocability. In addition, our data suggest that learners may not demonstrate a marked growth in backward collocability until they reach a more advanced level (i.e., B1_1 to B2+). It should be noted that this study focused on the quantitative analysis of L2 development in different aspects of collocation competence. We analyzed the text-based average scores of all the bigrams for different collocability metrics and therefore did not work on the analysis of collocation tokens that were specific to a particular syntactic schema. In other words, individual bigram tokens in each text may not be our major concern. The present study may be more helpful than Laufer and Waldman (2011) and Siyanova-Chanturia (2015) in that the developmental trends observed here can be more generalizable because they are based on the overall (or average) uses of all the two-word sequences in L2 texts.

The asymmetrical developments in forward-directed and backward-directed associations may have important implications for the development of L2 grammatical competence. Because learners write one word at a time during writing production, they would understandably have a stronger need to develop the competence in forward-directed word co-selection, while the acquisition of backward temporal relations is less pressing. Native speakers are normally able to make an intuitive judgement as to what the upcoming word should be given the preceding linguistic context, and learners need to acquire this skill for their L2 production. Our results have confirmed the significance of this forward collocation competence across learners of different proficiency levels. In our other project, we have also extended the analysis of directional collocability in two-word sequences to the lexical associations of multiword combinations beyond two-word collocations (Chen, 2019). A similar growth in forward-directed phraseological competence was also found in L2 uses of longer multiword combinations (cf. three- to five-word sequences in Chen [2019]). On the other hand, the late growth in backward-directed collocability is also found in Chen's (2019) analysis of multiword combinations beyond two-word sequences. Therefore, following Chen (2019), we posit that this lagging development of backward collocation competence may suggest a more sophisticated development in phrasal cohesiveness.

In recent years, a retrodiction-based learning has started to receive more attention in cognitive psychology. Humans can learn through both prediction-based (forward-directed) and retrodiction-based (backward-directed) association. A classic example was the experiment conducted in Jones and Pashler (2007), where human subjects could learn the varied forward and backward transitional probabilities of geometric shape sequences from the inputs and made correct predictions based on this implicitly-learned statistical knowledge. Moreover, the directionality of word-sequence associations may be related to the syntactic typology of the language in question. It has been found that language word order, or constituency structures, may act as a significant predictor of either higher forward or backward transitional probabilities in word sequences. For example, analyzing the English bigrams in the SUSANNE corpus, Onnis and Thiessen (2013) investigated the relationship between the bi-directional transitional probabilities of bigrams and the structures that these bigrams spanned. Their data suggest that bigram's backward transitional probability positively correlates with the phrase cohesiveness between the two words, i.e., tighter constituents in English formed by the two words. A bigram of high backward transitional probability is more often observed in words belonging to the same syntactic constituent or across a syntactic boundary that is at the lower syntactic level.

In particular, a backward-dominant association of a bigram W_1W_2 often implies that W_1 is limited in possibilities when W_2 is given, which can be connected to the head-initial right-branching structure of English. English syntactic maximal projections, such as a preposition phrase (PP), or complementizer phrase (CP), often take a functional head on the left and other lexical dependents on the right. This right-branching structure applies especially to common noun phrases and verb phrases where functional words like articles, determiners, or modals are positioned on the left-end of the phrase.

Let us illustrate the connection between phrasal cohesiveness and backward DP with some examples from L2 texts. For instance, in the prepositional phrase, *in advance*, the forward DP based on COCA is 0.0007, but its backward DP is 0.3365, almost five thousand times larger than the former. The asymmetrical strengths of the directional lexical associations suggest that given a content word like *advance, in* is one of the only few (functional) words that can precede it; however, many more words are likely to follow the preposition *in*. Similar asymmetrical backward-prominent lexical associations can also be found with bigrams that connect or mediate phrasally cohesive structures, such as noun phrases (e.g., *the importance*, *the impression*, *the happiness*), verbal phrases (e.g., *to abandon*, *to achieve*, *can contaminate*, *can extinguish*), and complementizer phrases (e.g., *that allows*, *that promotes*, *that connects*). The backward DP scores of the previous bigram examples are all stronger than their forward DP scores by at least three orders of magnitude. Figure 8 further provides the proportions of bigrams whose backward DP scores are larger than their forward DP scores by at least three orders of magnitude (i.e., backward DP/forward DP >= 1000) in terms of all bigrams whose first word is a functional word for each proficiency level. It is clear to see that the higher the learner proficiency, the more uses of backward-prominent bigrams. We argue that this may be preliminary evidence for advanced learners’ acquisition of L2 collocation competence at a phrasal (or grammatical) level—the increase in backward DP may indicate the L2 development of collocability in-between tighter constituents, thus leading to a higher level of phrasal cohesiveness in writing.

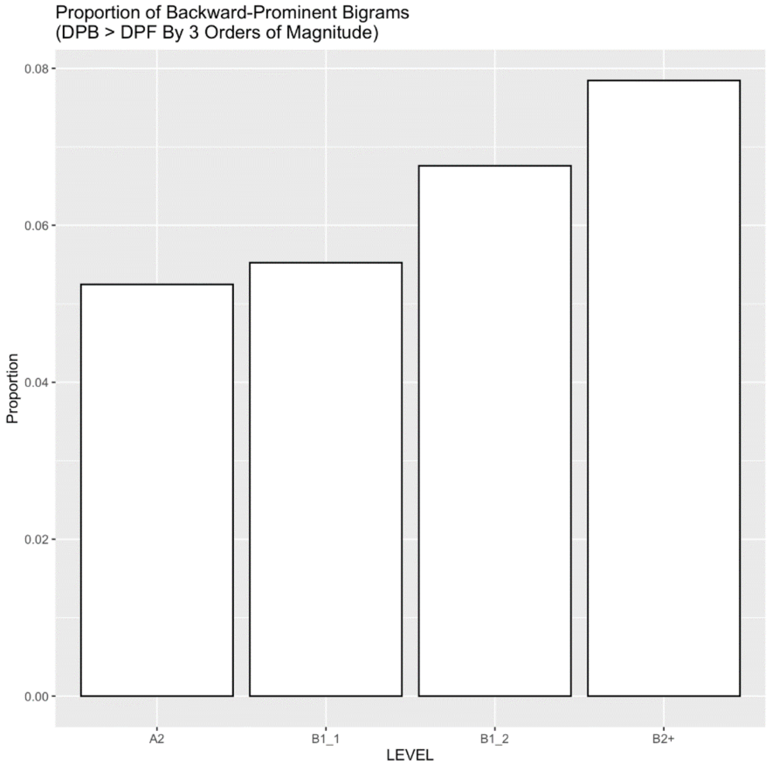

Figure 8. Proportion of backward-prominent bigrams by levels. Backward-prominent bigrams refer to the bigrams (a) whose first word is a functional word and (b) whose backward DP is larger than its forward DP by at least three orders of magnitude.

For dispersion, our analysis suggests that learners may start to use more domain-specific collocation patterns in the intermediate level (i.e., B1_1 to B1_2) because the IDF shows the most change on average in the transition of these two learning phases. Interestingly, on the other hand, the unseen rates show a more prominent decrease in the initial learning phases (i.e., from A2 to B1_2), suggesting that less proficient learners begin to use fewer bigrams that have not been used by native speakers when their proficiency progresses. These two findings both point to a developmental pattern in learners' collocation competence: as learners' proficiency grows, they start to use fewer deviant collocation patterns (i.e., unseen bigrams) but more domain-specific bundles (i.e., bigrams of high IDF).

CONCLUSION

This study has provided a more comprehensive analysis of the development of L2 collocation competence as learners grow in their proficiency. Different from previous research on phraseology, we examined the L2 collocation competence in a range of important dimensions of collocability. This study provides empirical evidence showing that learners use collocations that are more native-like in terms of exclusivity, directionality, and dispersion. Our findings are clear: learners do develop their collocation competence as their proficiency grows. Our analysis further suggests that learners develop this collocation competence more markedly, in terms of native-likeness, in the forward selection of words given the previous word. We suggest that a native-like intuition in backward selection of words may be developed in a more advanced learning phase. This may indicate that backward-directed collocation competence requires more implicit learning from extensive exposure to language input. Finally, our analysis of the dispersion and unseen rates has also highlighted a developmental pattern in learners' collocation competence: as learners' proficiency grows, they show a decreasing use of unseen bigrams, which are likely deviant collocation patterns, but an increasing use of domain-specific bundles. Overall, this study has provided a holistic account of the development of L2 collocational competence.

We would like to conclude this study by pointing out some of the directions for future research that stem from the limitations of the present study. The first limitation is concerned with the operational definition of learners' proficiency levels provided in the ICNALE. While proficiency levels of the learners were modeled and estimated based on well-received standardized English proficiency tests, a more rigorous validation may be needed to ensure the mapping between the test scores and CERF labels. Also, the test scores may represent a particular dimension of learner proficiency only. Secondly, this study is limited to two-word bundles that are adjacent to each other. Collocation competence may not necessarily be confined to contiguous two-word sequences (Bestgen, 2017; Gries, 2013).

For example, the Word Sketch Engine, a powerful on-line collocation toolkit, aims to capture collocations in a comprehensive range of long-distance grammatical relations (Kilgarriff et al., 2014). Although we have successfully extended the analysis of lexical associations to multiword combinations beyond bigrams (Chen, 2019), more research is needed to take into account the aspects of dispersion and creativity in multiword units.

Thirdly, this study only examines L2 essays in a particular genre, i.e., argumentative writings. Future studies are needed to examine the development of collocability competence in other contexts because studies have shown that phraseology varies considerably in different genres and/or registers, serving as effective linguistic scaffolding for creating domain-specific conventional texts (Biber et al., 2004; Hyland, 2008). Future studies may investigate how the development of L2 collocation competence may interact with these factors in a meaningful way. Another important issue that remains for further study is concerned with the fact that L2 collocation knowledge may be related to the structure of learners' L1 (Altenberg & Granger, 2001; Leńko-Szymańska, 2014). If the collocation patterns in L1 are similar to the patterns in L2, the development may be different. This may require an operational definition for cross-linguistic phraseological similarity. Finally, the present study has analyzed the development of the three aspects of collocability across proficiency levels independently. Future work is needed to further explore the inter-relationships among these three aspects, which may require a larger-scale analysis with more representative samples of each proficiency level.

REFERENCES

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes, 31*(2), 81–92.

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics, 22*(2), 173–194.

Appel, R., & Trofimovich, P. (2017). Transitional probability predicts native and non-native use of formulaic sequences. *International Journal of Applied Linguistics, 27*(1), 24–43.

Appel, R., & Wood, D. (2016). Recurrent word combinations in EAP test-taker writing: Differences between high- and low-proficiency levels. *Language Assessment Quarterly, 13*(1), 55–71.

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System, 69*, 65–78.

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.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgård & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 181–190). Amsterdam: Rodopi.

Biber, D., Conrad, S., & Cortes, V. (2004). *If you look at* ...: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.

Chen, A. C.-H. (2019). Assessing phraseological development in word sequences of variable lengths in second language texts using directional association measures. *Language Learning, 69*(2), 440–477.

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language, Learning & Technology, 14*(2), 30–49.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*(4), 397–423.

Crossley, S., & Salsbury, T. L. (2011). The development of lexical bundle accuracy and production in English second language speakers. *IRAL - International Review of Applied Linguistics in Language Teaching, 49*(1), 1–26.

Davies, M. (2012). *The Corpus of Contemporary American English (COCA): 560 million words, 1990-2012.* Retrieved from https://corpus.byu.edu/coca

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*(2), 157–177.

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics, 27*(1), 1–24.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014). The processing of verb-argument constructions is sensitive to form, function, frequency, contingency and prototypicality. *Cognitive Linguistics, 25*(1), 55–98.

Ellis, N. C., & Ogden, D. C. (2017). Thinking about multiword constructions: Usage-based approaches to acquisition and processing. *Topics in Cognitive Science, 9*(3), 604–620.

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

Evert, S. (2009). Corpora and collocations. In A. Ludeling & M. Kyto (Eds.), *Corpus linguistics: An international handbook* (pp. 1212–1248). Berlin: Mouton De Gruyter.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R.* Los Angeles: Sage Publications.

Firth, J. R. (1957). Modes of meaning. In J. R. Firth (Ed.), *Papers in linguistics 1934-1951* (pp. 190–215). Oxford: Oxford University Press.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning, 67*(1), 155–179.

Gries, S. T. (2010). Dispersions and adjusted frequencies in corpora: Further explorations. In S. T. Gries, S. Wulff, & M. Davies (Eds.), *Corpus linguistic applications: Current studies, new directions* (pp. 197–212). Amsterdam: Rodopi.

Gries, S. T. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics, 18*(1), 137–166.

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

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4–21.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world* (pp. 91–118). Kobe, Japan: Kobe University.

Jones, J., & Pashler, H. (2007). Is the mind inherently forward looking? Comparing prediction and retrodiction. *Psychonomic Bulletin & Review, 14*, 295–300.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography, 1*, 7–36.

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods, 50*, 1030–1046.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757–786.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), 647–672.

Leńko-Szymańska, A. (2014). The acquisition of formulaic language by EFL learners: A cross-sectional and cross-linguistic perspective. *International Journal of Corpus Linguistics, 19*(2), 225–251.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. *Cognition, 126*(2), 268–284.

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: Longman.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487–512.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Siyanova-Chanturia, A. (2015). Collocation in beginner learner writing: A longitudinal study. *System, 53*, 148–160.

Wood, D. (2015). *Fundamentals of formulaic language: An introduction*. London: Bloomsbury Publishing.

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

ACKNOWLEDGMENTS

The author would like to thank the anonymous reviewers of the *Taiwan Journal of TESOL* for their constructive comments to help improve earlier versions of this manuscript. This research was supported by a grant from the Taiwan Ministry of Science and Technology (108-2410-H-003-023-MY2).

CORRESPONDENCE

Alvin Cheng-Hsien Chen, Department of English, National Taiwan Normal University, Taipei, Taiwan.
Email address: alvinchen@ntnu.edu.tw

PUBLISHING RECORD

Manuscript received: May 11, 2020; Revision received: August 9, 2020; Manuscript accepted: August 24, 2020.