A longitudinal study of multi-word constructions in L2 academic writing: the effects of frequency and dispersion

Duygu Candarli[1] ![ORCID]

Abstract

This study examined the trajectories of the multi-word constructions (MWCs) in 98 advanced second language (L2) learners during their first-year at an English-medium university in a non-English-speaking country, using linear mixed-effects modelling, over one academic year. In addition, this study traced the academic reading input that L2 learners received at university, and it was investigated whether the frequency and dispersion of the MWCs in the input corpus would predict the frequencies of MWCs in L2 writers' essays. The findings revealed variations in the frequencies of different functional and structural categories of MWCs over time. This study provides empirical evidence for the effects of both frequency and dispersion of MWCs in the input corpus on the frequency of MWCs in L2 writers' essays, underscoring the importance of both frequency and dispersion in learning MWCs and the reciprocity of academic reading and writing. The findings have significant implications for usage-based approaches to language learning, modelling MWCs in L2 academic writing, and L2 materials design for teaching academic writing.

Keywords Multi-word constructions · Academic writing · Longitudinal study · L2 writing · Corpus linguistics

It is well-established that multi-word constructions (MWCs), such as 'on the other hand', constitute important discourse building blocks in English academic writing, mostly relying on noun and prepositional phrases (e.g., Biber, 2009; Biber, Conrad, & Cortes, 2004; Hyland, 2008). In the literature, a number of terms have been employed for MWCs, including 'lexical bundles' (Biber et al., 2004), 'formulaic sequences' (Wray, 2002), and 'academic formulas' (Simpson-Vlach & Ellis, 2010),

---

✉ Duygu Candarli
dcandarli001@dundeee.ac.uk

1 School of Education and Social Work, University of Dundee, Dundee DD1 4HN, UK

and these terms often have overlapping characteristics (Wray, 2002). The present study employs Liu's (2012) term 'multi-word constructions' to signal frequently occurring three-, four-, and five-word sequences in English academic writing and takes a usage-based approach to MWCs. As Liu (2012, p. 25) noted, "the term construction is adopted over expression/phrase/unit because it is a term preferred by contemporary linguistic theories such as Cognitive Linguistics."

Multi-word constructions have important discourse functions, as they introduce propositions (referential expressions), establish textual relations (discourse organisers), and express writers' (un)certainty and attitudes towards propositions (stance expressions) (Biber et al., 2004; Hyland, 2008). Referential expressions are the most common of the three discoursal categories of MWCs (operationalised as 'lexical bundles', i.e., frequently recurring word sequences) in published academic writing in English (Biber et al., 2004; Biber, 2009), followed by stance expressions. Discourse organisers are the least common discoursal category of MWCs in English academic writing. In terms of the structural categories, English academic writing relies on mostly noun phrases and prepositional phrases. Biber et al. (2004) note that there is an association between structural categories and discourse functions of MWCs in that referential expressions mostly consist of noun and prepositional phrases, whereas stance expressions are mostly verb phrases. Discourse organisers are composed of all structural types that include noun, prepositional and verb phrases (Biber et al., 2004). In L2 English writing, on the other hand, low-level writing in terms of the language proficiency relied on discourse organisers more than high-level writing, and high-level writers used referential expressions more frequently than low-level writers (Appel & Wood, 2016; Chen & Baker, 2016). While stance expressions were employed more frequently by high-level writers than low-level writers in Chen and Baker's study (2016), there was an opposite trend in Appel and Wood's study (2016). For L2 writers, the use of MWCs in academic writing is regarded as a marker of language proficiency, writing proficiency, and effective disciplinary communication (e.g., Hyland, 2008; Paquot, 2018; Paquot & Granger, 2012; Wray, 2002).

The number of L2 academic writers has been increasing as a result of the expansion of English-medium instruction at universities in non-English-speaking countries (Hyland, 2013). The first year at English-medium universities marks a key transitional period for L2 writers that learn how to write academic essays (Ortega & Iberri-Shea, 2005). Hence, there is a need to conduct longitudinal studies concerning the use of MWCs by L2 writers during their first year, in order to identify any changes, and to inform teaching practice at English-medium universities in non-English-speaking countries.

A growing body of research on MWCs inspired the development of empirically-derived lists of word combinations for academic writing, for use by researchers and teachers (e.g., Biber et al., 2004; Byrd & Coxhead, 2010; Liu, 2012; Simpson-Vlach & Ellis, 2010). Recent longitudinal studies on word combinations have revealed a complex picture of the development of phraseological performance of L2 learners in immersion settings in which the target language is spoken (e.g., Garner & Crossley, 2018; Kim, Crossley, & Kyle, 2018; Siyanova-Chanturia & Spina, 2020). However, less is known about trajectories of the MWCs employed by L2 writers over time at

English-medium universities in non-English speaking countries. In English-medium instruction contexts, students are taught all or most academic subjects in English, and they differ from English as a foreign language (EFL) contexts where classes other than English are taught in the learners' first language (L1). Also, no study has taken into account dispersion measures in the input corpus to research MWCs in L2 academic writing longitudinally. Building on the previous longitudinal research (e.g., Garner & Crossley, 2018; Siyanova-Chanturia & Spina, 2020) and filling these gaps, this study aimed to track the trajectories of the structural and discoursal categories of MWCs over one academic year at an English-medium university in Turkey. It also aimed to determine the effects of frequency and dispersion of MWCs in the input corpus on the frequency of MWCs in the L2 writers' essays, by tracing reading resources that students read at university. These reading materials were used as a proxy of their input for academic reading, since "reading and writing are reciprocal activities; the outcome of a reading activity can serve as the input for writing" (Grabe & Kaplan, 1996, p. 297).

Literature review

Frequency effects in longitudinal L2 studies

From a usage-based perspective, language is constituted of constructions that are "conventionalized pairings of form and function" (Goldberg, 2006, p. 3), and constructions range from a morpheme (e.g., 'ing'), to a word, and MWCs. Frequency of occurrence, which is regarded as the primary factor in language learning, is one of the major tenets of usage-based approaches to language (Ellis, 2002). Usage-based approaches to language hold that language learners are sensitive to "frequency, recency and context" (Ellis, 2002, p. 161–162) of constructions. The more frequently and recently language learners are exposed to a construction, the more frequently the construction is accessed and used (e.g., Ellis, 2002; Ellis, Römer, & O'Donnell, 2016). This suggests that the more frequent constructions are likely to be produced earlier by L2 learners than less frequent ones. Although token frequencies are an important aspect of learning constructions, other factors, such as salience, concreteness, also influence usage-based learning (Ellis, 2012).

In one of the earlier longitudinal studies on verb-argument constructions (VACs), such as 'verb + object + locative', Ellis and Ferreira-Junior (2009) found that seven L2 learners in Britain first used the most frequent verbs within each VAC, and that the frequencies of VACs used by L2 learners showed a significant correlation with those of VACs in the input of L1 speakers who were their conversation partners. Ellis and Ferreira-Junior (2009) acknowledged the limitations of using L1 data as a proxy of input for L2 speakers. More recent longitudinal studies that tracked lexical oral production by L2 learners (e.g., Crossley, Skalicky, Kyle, & Monteiro, 2019) found that L2 learners used more frequent words over a four-month period in the US, though lower level L2 learners used more infrequent words at the beginning than higher level learners. Both L1 and L2 token and type frequency norms were employed in Crossley et al.'s study (2019), and the frequency norms of TOEFL11

corpus were used as the proxy for L2 spoken frequency norms. Crossley et al. (2019, p. 17) argue that "early learners are not fully attentive to frequency distributions in the language, likely as a result of minimal exposure". These studies suggest that frequency of exposure influences L2 learners' use of constructions in longitudinal studies. Due to space limitations, this section is not exhaustive, and the frequency effects on the production of MWCs are reviewed later in the section on longitudinal studies.

Dispersion effects

Dispersion refers to the extent to which occurrences of a construction are (un) equally distributed in a corpus (Gries, 2008), and it can provide important information on how regularly learners are exposed to a construction. In addition to frequency values, information on dispersion is necessary because frequency alone might be misleading, since constructions that have similar frequencies in a corpus can be dispersed differently. Although there may be overall a negative correlation between frequency and dispersion of constructions, dispersion values of constructions within the middle frequency range can differ widely from each other (Gries, 2008). This suggests that both frequency and dispersion values of constructions need to be taken into account in learner corpus studies.

According to usage-based approaches to language learning, L2 learners retrieve and use constructions that not only occur frequently but also occur in a variety of contexts (e.g., Ellis, 2002; Ellis et al., 2016; Ellis & Wulff, 2015). As language users experience constructions more frequently and regularly, associative learning of constructions, which entails comparing and contrasting occurrences in context with previous ones, occurs over time, and this results in form-meaning mappings (Ellis et al., 2016). Ambridge, Theakston, Lieven and Tomasello (2006, p. 175) note that "learning is always better when exposures or training trials are distributed over several sessions than when they are massed into one session". The manifestation of 'distribution over several sessions' is dispersion (Gries, 2010). Similarly, Gablasova, Brezina and McEnery (2017, p. 160) regard dispersion as "an important predictor in language learning because collocations that occur across a variety of contexts are more likely to be encountered by language users".

Dispersion has received very little attention in empirical learner corpus studies. In a cross-sectional study that focused on L2 learners' lexical production (single words), the percentage of texts in which a word occurs, referred to as 'contextual diversity', was found to be the strongest predictor of the occurrences of verbs, while frequency effects were strongest for the occurrences of nouns in beginning-level L2 oral discourse at an American university (Crossley, Subtirelu, & Salsbury, 2013). It was also found that range, i.e., the number of texts in which a word occurs, referred to as 'contextual diversity', predicted the scoring of L2 essays better than token frequency in that L2 essays which included words used in constrained contexts received higher scores (e.g., Kyle & Crossley, 2016). It remains unknown to what extent dispersion of MWCs in the input corpus would predict the frequency of L2 learners' MWCs in academic writing in longitudinal studies.

Relationship between input and L2 writing

Usage-based approaches to language learning regard the linguistic input that learners receive as "the primary source for" L2 learning (Ellis & Wulff, 2015, p. 75). The constructions of the target language are learned through "experiencing their exemplars in contextualized usage" (Ellis et al., 2016, p. 42). Wulff (2019) argues that the concept of 'usage' not only involves learners' production but also their exposure to the target language. Hence, in this study, tracking L2 writers' academic essays and their compulsory academic reading resources as a proxy of their academic reading input longitudinally is theoretically motivated by usage-based approaches to language learning (Ellis et al., 2016; Ellis & Wulff, 2015) and reading-writing connections in the literature (see Grabe & Kaplan, 1996).

To date, no previous longitudinal study on the MWCs in L2 writing has tracked any reading input that learners receive. However, there are a few studies that made connections between the input learners received and linguistic features in L2 writing (e.g., Bi, 2020; Leedham & Cai, 2013). In a cross-sectional study, Leedham and Cai (2013) attributed Chinese students' frequent use of linking adverbials (e.g., *therefore*, *on the other hand*) in their academic writing to the teaching materials and textbooks that were used in China. In a recent study, informed by usage-based approaches, Bi (2020) examined linguistic features of argumentative and narrative essays of Chinese EFL learners and used reading passages in 16 textbooks as a proxy of their language input in order to investigate the effects of learners' input on their writing.

Review of longitudinal studies on multi-word constructions

Longitudinal studies of learner writing investigating the use of word combinations over time remain rare, although there has been "a slow and steady rise" in longitudinal learner corpus studies (Granger, Gilquin, & Meunier, 2016, p. 2). Early research used case or multiple-case study designs (e.g., Crossley & Salsbury, 2011; Li & Schmitt, 2009). For example, in a longitudinal case study of L2 academic writing in English, an MA student from an L1 Chinese background demonstrated a slow incremental learning of 'lexical phrases,' which were identified by three judges, over one academic year at a UK university (Li & Schmitt, 2009). Although the participant learned new phrases, the frequency or diversity of them showed no consistent patterns over time. Similarly, Crossley and Salsbury (2011) found an overall significant increase in the use of bigrams, which occurred frequently in the L1 reference corpus, in the speech of six L2 learners over one year in the US. However, not all the learners showed significant growth in the use of bigrams.

Longitudinal studies that included a larger number of participants have shown mixed results for the trajectories of word combinations in terms of frequency effects. Several studies revealed that fewer frequent combinations were employed by L2 learners over time (e.g., Bestgen & Granger, 2014; Siyanova-Chanturia & Spina, 2020). Bestgen and Granger (2014) examined 57 students' descriptive essays in

English in terms of bigrams at two time points over one semester at a US university and used a technique called 'CollGram', calculated through the *t*-scores and mutual information (MI) value and the proportion of absent bigrams in the reference corpus. These measures were based on the Contemporary Corpus of American English (COCA). The authors showed that though almost no change was observed in the collocations identified by the MI score, the students used gradually fewer high-frequency bigrams over one semester. Over the same period of time, Yoon (2016) explored verb–noun combinations in the argumentative and narrative essays of 51 L2 learners in the US at six different time points over one semester and found no significant changes in association strength of these combinations, derived from the COCA, in either genre over time. However, he found a difference between L1 and L2 writers' use of frequent collocations in that L1 writers used more infrequent collocations than their L2 counterparts. In a recent study, Siyanova-Chanturia and Spina (2020) analysed noun + adjective combinations in the essays of 175 L1 Chinese learners of Italian from beginner, elementary, and intermediate language proficiency levels, a larger sample than the previous studies, at two different time points over six months at a university in Italy. Using a large L1 Italian corpus as the benchmark, they reported that the L2 learners, irrespective of their proficiency level, used fewer frequent and strongly associated combinations over time. Hence, they concluded that the learners "started to use language more creatively and productively" (p. 28) as their exposure to L2 increased over time.

Other longitudinal research showed that L2 learners used more frequent combinations as a function of time (e.g., Garner & Crossley, 2018; Kim et al., 2018; Siyanova-Chanturia, 2015). Garner and Crossley (2018) examined the frequency, association, and proportion of bigrams and trigrams in the speech of 57 L1 Korean speakers of English in the US at four different time points over a four-month period and used the spoken section of the COCA to calculate *n*-gram indices. It was observed that changes in bigram frequency varied according to the speakers' proficiency levels in that the advanced learners employed more high-frequency bigrams at the initial status and that the high beginner students had the biggest growth in bigram frequency over time. The L2 speakers also produced a greater proportion of bigrams and trigrams that frequently occurred in L1 speech over time. Similarly, calculating *n*-gram indices based on the British National Corpus (BNC) and COCA, Kim et al. (2018) observed an increase in the use of more frequent bigrams and in the proportion of both bigrams and trigrams, which occurred in the L1 reference corpus, in the speech of six L2 learners in the US. In beginner-level L2 writing, it was found that the narrative essays of 36 L1 Chinese learners of Italian at a university in Italy included more frequent and more strongly associated noun + adjective combinations over five months, suggesting longitudinal improvement (Siyanova-Chanturia, 2015).

In a non-immersion setting, Zheng (2016) examined target-like lexical bundles that were identified using three empirically-derived lists in 15 first-year undergraduate academic essays of L1 Chinese writers in English and reported a U-shaped curve for the frequency of bundles during one academic year at a university in China. It was noted that participants' upper-intermediate proficiency and the duration of the study could account for the U-shaped curve in the use of lexical bundles. It should

be noted that abovementioned studies employed different methods to extract word combinations that were operationalised differently and included different L2 proficiency groups; therefore, direct comparisons between these studies are not possible. In longitudinal research on MWCs in L2 academic writing, the number of participants has ranged from 1 (Li & Schmitt, 2009) to 175 (Siyanova-Chanturia & Spina, 2020). Further longitudinal research is necessary to provide evidence of the trajectories of MWCs employed by L2 writers with different L1 backgrounds in under-researched contexts, specifically at English-medium universities in non-English-speaking counties.

The present study

Based on the review of studies above, the current study advances the longitudinal research on MWCs in L2 academic writing in several ways: First, most of the research reviewed above was conducted in immersion settings in which the target language was spoken. The context of this study was an English-medium university in a non-English-speaking country where students' exposure to English seemed more limited than immersion settings, but exposure to academic English could be richer at an English-medium university than in an EFL context. Second, most of the earlier studies investigated MWCs in L2 argumentative or narrative writing; therefore, this study extends the analysis of MWCs to discipline-specific academic writing longitudinally. The focus on disciplinary academic writing is important because degree programmes at English-medium universities require L2 writers to write discipline-specific academic writing (e.g., Hyland, 2013). Also, MWCs (operationalised as lexical bundles in these studies) vary across different disciplines in published academic writing (e.g., Cortes, 2004; Hyland, 2008) and in university student's writing (e.g., Durrant, 2017) in terms of their discourse functions. Third, this study traced the reading resources that students were asked to read as part of their university courses, and these were conceptualised as a proxy of their academic reading input. This made it possible to investigate to what the extent the frequency of MWCs in L2 writing is tuned by 'usage' experiences, as operationalised by the frequency and dispersion of MWCs in the academic readings of L2 writers, further exploring the relationship between time and the frequency and dispersion of MWCs in the academic reading input corpus. Hence, the present study responds to the call for integration of 'usage' experiences (Ellis et al., 2016; Wulff, 2019) in learner corpus studies. Although no corpus can fully reflect each individual student's language experiences, this approach would contribute to ecological validity, which is "the degree of similarity between a research study and the authentic context that the study is purportedly investigating" (Loewen & Plonsky, 2016, p. 56). MWCs vary across different genres (academic, fiction, etc.) (Biber et al., 2004; Hyland, 2008); therefore, using L1 general corpus or any other L2 corpus as a proxy of the input may not capture L2 learners' academic reading experiences.

Fourth, dispersion, an important predictor of L2 learning (Gries, 2010), has yet to be investigated in longitudinal research on MWCs. This study used Gries' deviance of proportions (described in the methods section), a more fine-grained measure of

dispersion than range (Gries, 2008), to analyse to what extent the dispersion of the MWCs in the reading resources (input) would predict the frequency of MWCs in L2 writers' essays. Fifth, most of the previous longitudinal research looked at MWCs in terms of different lengths (e.g., bigrams, trigrams) or their part-of-speech tags (e.g., noun + adjective combinations). This study operationalised MWCs at six different levels, using Biber et al.'s (2004) taxonomy: noun phrase-based MWCs, prepositional phrase-based MWCs, and verb phrase-based MWCs in terms of their structural categories, and referential expressions, discourse organisers, and stance expressions in terms of their discoursal categories. It also examined a range of sequences, namely three-, four- and five-word MWCs. Finally, in the context of a 'multilingual turn' (Ortega, 2018, p. 65) that recognises the value of researching L2 learners' longitudinal language use on its own rather than comparing it with L1 benchmarks, this study used no L1 reference corpus. The present study is exploratory in nature and addresses the following research questions:

1. To what extent, if any, does the frequency of MWCs change in the essays of L2 writers, in terms of structural categories and discourse functions, over one academic year?
2. To what extent, if any, do the frequency and dispersion of MWCs in the academic readings of L2 writers predict the frequency of MWCs in their essays over time?

Methods

Participants

The 98 participants of this study were first-year university students on an English Language Education programme at an English-medium university in Turkey. At the outset of the academic year, they were given a participant information sheet, providing information about the research study. The participant information sheet included no specific examples for MWCs. The students gave the researcher informed written consent to use their academic assignments during their first year. The participants also completed a questionnaire requesting information concerning their first language, gender, previous residency in an English-speaking country, proficiency in other languages, and the medium of instruction at their secondary school. The first language of all the participants was Turkish, and they were aged between 17 and 22 years ($M = 18.41$, $SD = 1.16$). The majority of the participants (83%, $n = 81$) were female, and 17% ($n = 17$) were male. None of the participants had resided in an English-speaking country for more than one month, or possessed advanced language proficiency in another language. For all the participants, the medium of instruction at secondary school was Turkish. Throughout their undergraduate education at the English-medium university, they submitted their assignments in English, except two course units that were in Turkish. The participants had four compulsory course units per semester, and the classroom contact time varied between two and three hours per course unit, per week. They took an 'Academic Writing' and 'Study and Research Skills' course units in the first and second semester of their first year, respectively,

and they read resources on academic writing processes and strategies and analytical academic writing. There was no explicit teaching of MWCs in academic writing class, as reported by the lecturers and students in the interviews (Candarli, 2020). The course units focused on developing essay structure, analysis and synthesis of academic sources, paraphrasing, and citation conventions.

Before commencing their studies, the students were required to pass the university's English proficiency test with a good score, the equivalent to an overall band of 6.5 in IELTS (Academic), with no less than 6.5 in writing, or TOEFL IBT (at least 79). These minimum scores correspond to a borderline B2/C1 level of the Common European Framework of Reference (CEFR) of Languages (Taylor, 2004). The scores of the students ranged from those that were equivalent to 6.5 to 7.5 in IELTS, $M = 6.63$, $SD = 0.28$, and most of the students scored 6.5 in IELTS Academic or in an equivalent test. Hence, the participants can be regarded as advanced L2 learners. Since these students attended secondary schools at which the medium of instruction was Turkish, their first year at an English-medium university represented a transitional stage from secondary school to university.

Essays

This study used L1 Turkish university students' academic assignments submitted for their assessed course work at an English-medium university in Turkey. These assignments were collected from the same participants at three stages during one academic year: The beginning of November (Month 3), the end of January (Month 5), and the beginning of June (Month 9). The assignments, which all received passing grades, at least 50 out of 100 at university, were checked for plagiarism, using 'Turnitin' to which the students submitted their assignments. The discipline-specific, un-timed written assignments featured similar topics, including gender differences in education and social media use in education (see S1 in supplementary material for the assignment prompts). The participants were free to consult any reference materials while writing, but no data were collected on this, since source text use was beyond the scope of this study. The students wrote these essays for their assignments in their academic subject rather than for research purposes, which increased the ecological validity of this study. It is worth noting that the researcher was not a lecturer at the university where data were collected and had no control over the topics of the assignments or reading lists of the L2 writers at the university. These assignments can be regarded as 'analytical exposition' (Coffin, 1996) essays, which require students to engage with the extant literature, to evaluate and synthesise the arguments therein, and to present their own position. Analytical exposition essays fall within the 'essay' genre family in terms of Nesi and Gardner's (2012) taxonomy of the genres of discipline-specific student writing in UK higher education, and they constitute a hybrid genre of 'exposition' and 'discussion'. As in Li and Schmitt's (2009) study, the list of references and direct quotations were removed from all the essays.

As Table 1 illustrates, the number of tokens in the L2 writers' essays increased over time, especially at Month 9 because the suggested word limit was 1500 for the final assignment instead of 500 words at Month 3 and 5. Eight of the students'

Table 1 The characteristics of the essays

	Month 3	Month 5	Month 9
Number of texts	98	98	90
Number of tokens	50,703	56,477	160,013
Average text length and standard deviation (SD)	517 (103)	576 (115)	1778 (438)
Minimum–maximum text length	302–793	327–815	719–2945

essays were absent at Month 9, since they did not submit an essay in June, due to a variety of reasons, including dropping out of university, and mitigating circumstances that enabled submission at a later time.

Input corpus

Input corpus included compulsory readings of the compulsory modules that the participants of this study, one cohort of first-year university students, took during their first year at university. This corpus was built in order to determine whether the frequency and dispersion of MWCs in the reading materials that the students encountered would predict the frequency of MWCs in the L2 writers' essays (see S3 in supplementary material for the reading resources). Although these academic texts were arguably an estimation of students' academic reading input and a potential source of target-like academic MWCs, this study does not argue that they constituted the only input for students. The participants took lectures in English, and it is likely that they read English materials and watched television series in English in their free time; however, it is not possible to capture all the input that students were exposed to. Within the usage-based approaches to language learning, input is mainly operationalised in two ways: (1) An L1 corpus (e.g., Ellis & Ferreira-Junior, 2009); (2) Textbooks that students use for their classes (e.g., Bi, 2020). Given that "English neither functions for intranational communication purposes, nor is used for basic communicational goals in Turkish society" (Selvi, 2020, p. 4), the course readings of the L2 writers at university served as the main input in the context of this study (see Bi, 2020). It was also ecologically more valid to determine what kind of input L2 learners were exposed to in non-immersion settings, since "frequency in a general corpus, even one constructed from second language learner speech, is not necessarily the frequency with which a particular learner experiences the form" (Larsen-Freeman, 2015, p. 238).

The reading materials included mostly book chapters and a research article, and the soft copy versions of these resources were obtained as much as possible. When a soft copy was not available, book chapters were scanned, and optical character recognition (OCR) was applied to convert these to plain text files, using the tesseract package (Ooms, 2018) in R (R Core Team, 2019). Any OCR errors were checked and corrected, using Notepad + +, a text editor. All reference lists were removed from each individual file. As seen in Table 2, the participants of this study were assigned to read 74 texts by Month 9 (June) for their compulsory modules, and each

Table 2 The characteristics of the input corpus

	By Month 3	By Month 5	By Month 9
Number of texts	22	35	74
Number of tokens	321,563	549,015	956,091
Average text length and standard deviation (SD)	14,617 (6771)	15,686 (7373)	12,920 (6587)
Minimum–maximum text length	4556–27,623	4556–28,479	4391–28,964

reading source in the list (a book chapter or a journal article) was operationalised as a text. There were 22 individual texts by Month 3, 35 individual texts (22 + 13) by Month 5 and 74 texts (22 + 13 + 39) by Month 9, which reflected L2 writers' cumulative exposure to academic English during their first year, and this corpus was one source of input that was used as a proxy of the L2 writers' academic reading input. It is worth noting that the students may or may not have referred to these sources at the time of writing their assignments. Due to the laborious nature of scanning book chapters and checking OCR errors, the input corpus only included texts that were assigned as 'compulsory' in the reading lists of the participants' compulsory modules during their first year.

Identification of MWCs

The MWCs in the L2 writers' essays and in the input corpus were identified using three empirically-derived lists of MWCs (see S2 in supplementary material for the list of MWCs): (a) Biber et al.'s (2004) list of lexical bundles (four-word sequences) that occurred at least 10 times per million words in academic prose; (b) Liu's (2012) list of the most frequently used MWCs in academic writing (excluding two-word sequences); (c) Simpson-Vlach and Ellis' (2010) written academic formulas (three- to five-word sequences). Despite different terms (lexical bundles, MWCs, and academic formulas) were used in these studies, they all referred to multi-word sequences that have certain discourse functions in context and occur frequently in academic writing. The three lists were used to identify MWCs in this study for three reasons. First, the corpus of L2 writers' essays in this study was too small to extract MWCs from the corpus itself, especially at Month 3 and Month 5, since Cortes (2013) argued that a corpus consisting of at least one million words is required to extract lexical bundles from the corpus itself. Second, the MWCs in these lists could minimise topic effects (Yoon, 2016), since they were extracted from large corpora and not topic-bound sequences. Third, the MWCs in the lists served as a proxy for target-like academic MWCs, since the frequency of occurrence and range in academic prose were identification criteria for the MWCs in these lists. In order not to inflate token frequencies, the lists of MWCs were adapted in several ways. First, in the cases of overlaps of MWCs of different lengths featured in the lists, such as 'as well as' and 'as well as the', only the shorter MWC was counted in the corpora, and longer ones were removed from the lists, except in the case of 'on the other hand' which was selected instead of 'on the other'. When there were partial overlaps of

MWCs, such as 'more likely to' and 'is more likely', the concordance lines were checked to see whether they occurred within the same co-text, and the token frequencies were noted accordingly. For example, when 'more likely to' occurred 30 times in the corpus and 'is more likely' occurred seven times, the occurrence of 'is more likely to' ($n=3$) was checked. Then, the frequency of 'is more likely to' ($n=3$) was subtracted from the frequency of 'is more likely' ($n=7$) to record the frequency of 'is more likely' (see Chen & Baker, 2016). Place names, such as 'in the United States' were excluded from the list of MWCs. Lastly, in Liu's (2012) list of MWCs, the constructions of two words with a schematic representation, such as 'NP suggest that' and 'according to (det+N)' were excluded. All the MWCs that were compiled from the abovementioned empirically-derived lists were searched in both the L2 writers' essays and input corpus, using a free corpus tool, #LancsBox version 3.03 (Brezina, McEnery, & Wattam, 2015). Within #LancsBox, the Whelk tool provided frequencies of each MWC for each text. Then, the frequencies of each MWC were recorded on a spreadsheet for each text.

Analysis of MWCs

In terms of analysis, all the MWCs identified in both the L2 writers' essays and the input corpus were coded structurally, employing the taxonomy of previous studies (Biber et al., 2004; Chen & Baker, 2016), as shown in Table 3. Table 3 also shows the different number of structural types of MWCs that occurred in the L2 writers' essays over time. Due to the different size of the corpora of the L2 writers' essays, only the token frequencies of the MWCs were investigated in this study. This also applies to the discoursal categories of MWCs.

The MWCs were also coded according to their discourse functions in both the L2 writers' essays and in the input corpus. Several taxonomies have been proposed for the discourse functions of MWCs (e.g., Biber et al., 2004; Hyland, 2008). This study employed an adaptation of Biber et al.'s (2004) taxonomy of the discourse functions of lexical bundles for two reasons: First, it is widely used in the literature of academic discourse (Cortes, 2013). Second, Simpson-Vlach and Ellis' (2010) classification scheme and Liu's (2012) semantic functional categories of MWCs draw on and show similarities with Biber et al.'s (2004) taxonomy of functional categories. Biber et al. (2004) classified lexical bundles into three main categories: (a) referential expressions, which introduce abstract and concrete entities, and frame propositions; (b) discourse organisers, which signal causative, inferential, and transitive relations in a text; (c) stance expressions, which convey the (un)certainty of the writer, express the writer's attitudes, and indicate obligations or ability. Biber et al.'s taxonomy (2004) was adapted in two ways. 'Descriptive' MWCs (Cortes, 2004) that indicate abstract and concrete entities (e.g., 'the concept of') were added to the main category of 'referential expressions'. 'Inferential/resultative signals' were added to the main category of 'discourse organisers' to indicate cause-effect relations (e.g., 'as a result') in a text (Hyland, 2008).

All the MWCs were coded according to the functional taxonomy presented in Table 4, by examining the concordance lines and wider co-text of each MWC in

Table 3 Structural categories of MWCs and different number of types that occurred in the L2 writers' essays

Structural categories of MWCs	Example MWCs	Month 3		Month 5		Month 9	
		3-word MWCs	4-word MWCs	3-word MWCs	4-word MWCs	3-word MWCs	4-word MWCs
1. Noun phrase-based (NP-based) MWCs							
1.a. Noun phrase with of-phrase fragment	the purpose of	67	5	61	4	74	7
1.b. Noun phrase with other post-modifier fragments	important role in						
2. Prepositional phrase-based (PP-based) MWCs							
2.a. Prepositional phrase with embedded of-phrase fragment	in terms of	19	5	19	7	25	6
2.b. Other prepositional phrases	at the same time						
3. Verb phrase-based (VP-based) MWCs							
3.a. (Noun phrase+) copula be+noun phrase/adjectival phrase	is one of	37	8	45	7	67	13
3.b. Verb phrase with active verb	make use of						
3.c. Anticipatory it+verb phrase/adjectival phrase+(complement clause)/to-clause fragment	it is possible to						
3.d. Passive verb+(prepositional phrase fragment)	can be seen						

Table 4 The functional taxonomy of discourse functions of MWCs and different number of types that occurred in the L2 writers' essays

Discourse functions of MWCs		Example MWCs	Month 3			Month 5			Month 9		
			3-word MWCs	4-word MWCs	5-word MWCs	3-word MWCs	4-word MWCs	5-word MWCs	3-word MWCs	4-word MWCs	5-word MWCs
1. Referential expressions	1.a. Identification/focus	one of the most	94	8	–	90	8	–	103	12	–
	1.b. Specifying attributes (quantifying/ framing)	in this case									
	1.c. Descriptive	the concept of									
	1.d. Place-/time-/ text-deictic	the beginning of									
2. Discourse organisers	2.a. Topic introduction/focus	in this paper	10	2	1	9	2	1	18	2	1
	2.b. Topic elaboration/clarification/transition	in addition to									
	2.c. Inferential/ resultative signals	as a result									
3. Stance expressions	3.a. Epistemic stance	it seems that	24	8	–	26	6	–	54	10	–
	3.b. Attitudinal stance/modality	it is important									

Word Smith Tools 6.0 (Scott, 2012). Table 4 also shows the different number of discoursal types of MWCs that occurred in the L2 writers' essays over time. When an MWC possessed multiple discourse functions, the predominant function of each MWC in the data was coded as the functional category (e.g., Chen & Baker, 2016). In order to assess inter-coder agreement, about 25% of the MWCs ($n = 59$) identified at Month 9 were coded separately by another researcher in applied linguistics, and the Cohen's kappa value was 0.90, which indicated "almost perfect agreement" according to Landis and Koch's guidelines (1977). After that, the differences were resolved through discussion. The MWCs that did not fit into categories of structural or discoursal categories of MWCs were coded as 'others' and excluded from further analysis.

In addition to the frequency analysis of structural and discoursal categories of MWCs, dispersion measure of MWCs was calculated in the input corpus in order to investigate whether dispersion of MWCs in the reading materials would predict their frequency in the L2 writer's essays. As a dispersion measure, Gries' (2008) (normalised) deviance of proportions (DPnorm), which was refined in Lijffijt and Gries (2012), was calculated since DPnorm can handle differently-sized corpus parts and provide a value between 0 and 1, which is easy to compare across studies. Each book chapter or journal article in the reading lists was a corpus part in this study. DPnorm was calculated in the following way (Gries, 2008; Lijffijt & Gries, 2012): (1) the size of each corpus part was computed as percentages of the whole corpus, and expected percentages of a MWC were determined; (2) token frequencies of a MWC within each corpus part were calculated as observed percentages; (3) the absolute pairwise differences between (1) and (2) were computed, summed up and divided by 2. DPnorm can take a value between 0 and 1, which means even and uneven dispersion, respectively. In this study, dispersion was operationalised as the normalised dispersion of MWCs in the input corpus, while frequency was operationalised as the normalised token frequencies of MWCs.

Statistical analyses

In this study, linear mixed-effects models (LMMs) were employed to analyse changes in the frequencies of MWCs. Mixed/mixed-effects models were preferred over traditional ANOVA, since mixed-effects models quantify both group-level and individual-level patterns within a single analysis, taking into account sources of random variation (e.g., Gries, 2015; Linck & Cunnings, 2015; Murakami, 2016). Mixed-effects models are also robust enough to handle missing data (Linck & Cunnings, 2015).

In order to answer the first research question, individual essays served as the unit of analysis. The frequencies of each main structural category of MWCs (NP-based MWCs, PP-based MWCs, and VP-based MWCs), and each main discoursal category of MWCs (referential expressions, discourse organisers, and stance expressions) were recorded for each essay, and were then normalised per 500 words per text (each text received a normalized, per 500 words, frequency count for each category of MWCs). The recording of the frequencies for each text in a learner corpus would enable

generalisations about learners' language systems (see Durrant & Schmitt, 2009). Two separate LMMs were built to depict the trajectories of the structural categories (model 1) and discoursal categories (model 2) of MWCs. There was not enough data to build models for subcategories of structural and functional categories of MWCs. The dependent variable was the normalised frequency of each category of MWCs for both models (a unique dependent variable for each category of MWCs at each time point in the long data format). Time (months in academic year—3, 5, and 9—categorical variable) was added as a fixed effect. The variables 'structural_category' (NP-based MWCs, PP-based MWCs, VP-based MWCs—categorical) and 'discoursal_category' (Referential expressions, discourse organisers, stance expressions—categorical) were added as the second fixed effects for the model for the structural categories and discoursal categories of MWCs, respectively. The L2 writers' English proficiency test scores were also added as the fixed effects variables for both models. The random effects, i.e., those that account for individual variation, were L2 writers with random intercepts and slopes of time and 'structural_category' (model 1)/ 'discoursal_category' (model 2) and their interactions for writers. All the models in this study were fit with *lme4* package version 1.1–21, using *lmer* function (Bates, Mächler, Bolker, & Walker, 2015) in R version 3.6 (R Core Team, 2019). Then, post-hoc tests, using the Tukey adjustment, were conducted to estimate changes in each category of MWCs across time in *lsmeans* package (Lenth, 2016). The next section will only present the post hoc tests that estimate longitudinal trajectories of each category MWCs rather than pairwise differences between different categories of MWCs at each time point.

In this study, the text length and the frequencies of MWCs were missing for eight students out of 98 at Month 9 because students had dropped out of university or submitted their essays at a later time. Out of 882 data points in the long data format, only 2.7% of the frequency ($n=24$) and 2.7% of ($n=24$) text length data points were missing for structural categories and discoursal categories of MWCs, respectively. Since Schafer (1999) argued that missing data points of 5% or less are inconsequential, all the data were included in the models which discarded the missing data points at only Month 9 rather than all the data points of a learner.

Individual MWCs served as the unit of analysis in order to address the second research question. Two separate LMMs were built to determine to what extent time, frequency and dispersion of MWCs in the input data would predict frequencies of the structural categories (model 3) and discoursal categories (model 4) of MWCs in the L2 writers' essays. The dependent variable was the normalised frequency of each MWC (per 500 words per text) for both models in the L2 writers' essays. The fixed effects variables were as follows: (1) Normalised frequency of MWCs in the input corpus (per 500 words per corpus); (2) time (months in academic year—3, 5, and 9—categorical variable); (3) DPnorm (normalised DP values for MWCs in the input corpus); (4) Scores of the L2 writers' English proficiency tests; (5) The variables 'structural_category' (NP-based MWCs, PP-based MWCs, VP-based MWCs—categorical) and 'discoursal_category' (Referential expressions, discourse organisers, stance expressions—categorical) for the model for the structural categories and discoursal categories of MWCs, respectively. MWCs and L2 writers were included as crossed random effects (see Gries, 2015). The random effects structures at first involved random intercepts and slopes of time and 'structural_category' (model 3)/

'discoursal_category' (model 4) and interactions for both MWCs and writers. The random effects structures had to be simplified for both models due to the model convergence issues even with optimisers (see Bates et al., 2015).

For all the four models, optimal random effect structures were selected first, and then optimal fixed effect structures (see Durrant & Brenchley, 2019; Gries, 2015). In order to achieve this, Akaike information criterion (AIC), which provides a relative goodness of fit of different models, was used. The smaller the AIC value, the better the fit the model provides for the data (Murakami, 2016). In terms of model selection, the backward selection heuristic, commencing with the most complex model, with both fixed (all possible fixed effects and their interactions) and random effects that involved maximal random effects structures (random intercepts and slopes for all possible predictors and their interactions—maximal random effects structure for model 1 and 2) was followed (Barr, Levy, Scheepers, & Tily, 2013). The model complexity was reduced until a further reduction indicated a bigger AIC value (Murakami, 2016). *P* values were derived from the models, using *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017). The effect sizes were calculated using *MuMIn* package version 1.43.15 (Bartoń, 2019). Significance of random effects was evaluated, using a parametric bootstrap test with *pbkrtest* package (Halekoh & Højsgaard, 2014) because parametric bootstrapping could provide more accurate results than the likelihood ratio test (e.g., Bates et al., 2015). Finally, the models met the assumptions of mixed-effects models (see Durrant & Brenchley, 2019) with regard to the normal distribution of residuals and random effects, linear relationship between residuals and predicted values, and homogeneity of residual variance. These were checked via plots with *performance* package (Lüdecke, Makowski, & Waggoner, 2019) in R. Also, no multicollinearity was found between predictors; and there were no outliers.

Results

This section addresses the first research question on the changes in the frequency of MWCs in the L2 writers' essays across their structural and discoursal categories and then the second research question on the effects of frequency and dispersion of MWCs in the readings of L2 writers on the frequency of MWCs in the L2 writers' essays. Descriptive statistics for the normalised frequencies of each structural category of MWCs and for the normalised frequencies of each discoursal category of MWCs are presented in Table 5 and in Table 6, respectively.

Table 5 Descriptive statistics for the frequencies of structural categories of MWCs

	Month 3	Month 5	Month 9
NP-based MWCs	1.88 (1.49)*	2.09 (1.82)	2.42 (1.09)
PP-based MWCs	1.69 (1.17)	1.51 (1.00)	1.44 (1.00)
VP-based MWCs	1.29 (1.19)	1.45 (1.24)	1.56 (0.88)

*Mean (Standard deviation)

Table 6 Descriptive statistics for the frequencies of discoursal categories of MWCs

	Month 3	Month 5	Month 9
Referential expressions	2.34 (1.98)	2.58 (1.77)	2.92 (1.47)
Discourse organisers	1.65 (1.44)	1.44 (1.24)	1.22 (1.08)
Stance expressions	0.86 (1.09)	1.02 (1.02)	1.29 (0.96)

*Mean (Standard deviation)

The trajectories of structural categories of MWCs

There was a significant effect of 'Month 9' on the frequency of NP-based MWCs (e.g., *the use of*), indicating that their frequency increased at Month 9 in comparison to Month 3 in the L2 writers' essays, as seen in Table 7. From Month 5 to Month 9, there was no significant change in the frequency of NP-based MWCs ($t=-1.83$, $p=0.16$), as post-hoc tests showed. A significant interaction of 'Month 9' and PP-based MWCs was observed. However, post-hoc tests showed no significant change in the frequency of PP-based MWCs (e.g., *in terms of*) from Month 3 to Month 9 ($t=1.54$, $p=0.27$), or from Month 3 to 5 ($t=1.14$, $p=0.49$), or from Month 5 to 9 ($t=0.43$, $p=0.90$). There was no significant interaction between 'Month 9' and VP-based MWCs, suggesting that time had no significant effect on the frequency of VP-based MWCs in the L2 writers' essays. Post-hoc tests revealed that the frequency of VP-based MWCs (e.g., *I argue that*) showed no significant change from Month 3 to Month 9 ($t=-1.57$, $p=0.26$), or from Month 3 to 5 ($t=-0.99$, $p=0.58$) or from Month 5 to 9 ($t=-0.59$, $p=0.82$). None of the interactions between 'Month 5'

Table 7 Mixed-effects model for the structural categories of MWCs

Parameters	Fixed effects			
	Estimate	*SE*	*t*	*p*
(Intercept)*	1.88	0.14	13.54	<0.001
Month 5	0.21	0.16	1.31	0.19
Month 9	0.52	0.17	3.10	0.002
PP-based MWCs	−0.19	0.17	−1.08	0.28
VP-based MWCs	−0.59	0.18	−3.32	<0.001
Month 5: PP-based MWCs	−0.40	0.23	−1.72	0.08
Month 9: PP-based MWCs	−0.78	0.24	−3.29	0.001
Month 5: VP-based MWCs	−0.05	0.23	−0.22	0.83
Month 9: VP-based MWCs	−0.26	0.24	−1.09	0.28
	Random effects			
	Variance	*SD*	Corr	
(Intercept)|ID	0.58	0.76		
PP-based MWCs|ID	0.30	0.55	−0.68	
VP-based MWCs|ID	0.47	0.69	−0.57	

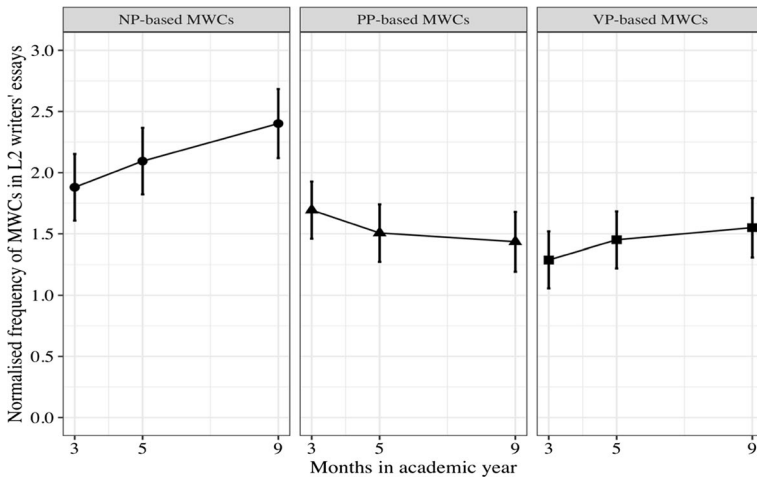*Reference level (intercept) is NP-based MWCs at Month 3

Fig. 1 Predicted values for the frequencies of the structural categories of MWCs (vertical lines indicate 95% CI)

and any structural categories of MWCs were significant. The variable 'proficiency scores' was dropped from the model, since it did not improve the model fit, according to AIC values. Figure 1 shows the predicted frequencies of the structural categories of MWCs.

Table 7 shows that the random effects were negatively correlated in that the L2 writers' essays which included more PP-based MWCs at Month 3 had a more rapid decrease in the frequency of PP-based MWCs over time. Similarly, the L2 writers who used fewer VP-based MWCs at Month 3 increased their use of VP-based MWCs more rapidly. R^2 marginal of the model, which shows variance explained by the fixed effects alone, was 0.16, indicating limited predictive power. R^2 conditional of the model, which indicates variance explained by the whole model (fixed and random effects), was 0.22. A parametric bootstrap test indicated that random slopes were statistically significant ($p=0.01$), whereas random intercepts were non-significant ($p=0.99$).

The trajectories of discoursal categories of MWCs

The frequency of referential expressions (e.g., *one of the most*) increased at Month 9 in the L2 writers' essays in comparison with Month 3, as shown in Table 8. Post-hoc tests indicated no significant change in the frequency of referential expressions from Month 5 to Month 9 ($t=-1.70$, $p=0.21$). A significant interaction of 'Month 9' and discourse organisers (e.g., *on the other hand*) was found, and post hoc tests revealed that there was a significant decrease in the frequency of discourse organisers from Month 3 to Month 9 ($t=2.37$, $p=0.04$). No significant change occurred in the frequency of discourse organisers from Month 3 to Month 5 ($t=1.10$, $p=0.51$) or from Month 5 to Month 9 ($t=1.30$, $p=0.40$). No significant interaction was found between

Table 8 Mixed-effects model for the discoursal categories of MWCs

Parameters	Fixed effects			
	Estimate	SE	t	p
(Intercept)*	2.34	0.15	15.59	<0.001
Month 5	0.24	0.18	1.29	0.20
Month 9	0.56	0.19	2.96	0.003
Discourse organisers	−0.69	0.19	−3.68	<0.001
Stance expressions	−1.48	0.20	−7.51	<0.001
Month 5: Discourse organisers	−0.44	0.26	−1.69	0.09
Month 9: Discourse organisers	−1.00	0.27	−3.78	<0.001
Month 5: Stance expressions	−0.07	0.26	−0.27	0.79
Month 9: Stance expressions	−0.12	0.27	−0.46	0.65
	Random effects			
	Variance		SD	Corr
(Intercept)|ID	0.56		0.75	
Discourse organisers|ID	0.19		0.44	−0.56
Stance expressions|ID	0.54		0.73	−0.68

*Reference level (intercept) is referential expressions at Month 3

'Month 9' and stance expressions (e.g., *are likely to*), meaning that the frequency of stance expressions showed no change as a function of time. Post hoc tests showed no significant change in the frequency of stance expressions from Month 3 to Month 9 ($t = -2.31$, $p = 0.06$), or from Month 3 to Month 5 ($t = -0.92$, $p = 0.63$), or from Month 5 to Month 9 ($t = -1.42$, $p = 0.33$) in the L2 writers' essays. None of the interactions between 'Month 5' and any discoursal categories of MWCs were significant, as seen in Table 8. The variable 'proficiency scores' was dropped from the model, since it did not improve the model fit, according to AIC values. Figure 2 shows the predicted frequencies of the discoursal categories of MWCs over time.

As shown in Table 8, negatively correlated random effects indicated that the L2 writers' essays with more frequent discourse organisers at Month 3 had a more rapid decline in the frequency of discourse organisers over time. On the other hand, the L2 writers who used fewer stance expressions increased their use of stance expressions more rapidly over time. R^2 marginal of the model was 0.20; and R^2 conditional of the model was 0.32. A parametric bootstrap test indicated that random slopes for the L2 writers were statistically significant ($p = 0.02$), whereas random intercepts were non-significant ($p = 0.74$).

Effects of frequency and dispersion of MWCs in the input corpus on the frequencies of MWCs in the L2 writers' essays

Structural categories of MWCs. Descriptive statistics for the normalised frequency and DPnorm of MWCs are shown in Table 9. It is worth noting that the normalised frequency figures of MWCs are very low because the analysis was conducted at the
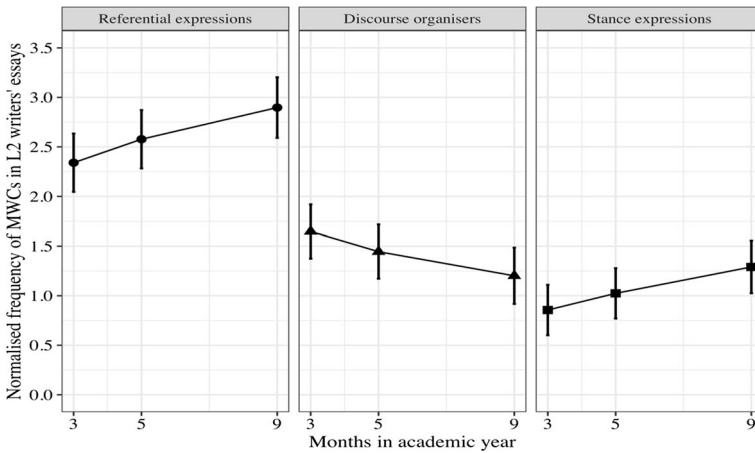
Fig. 2 Predicted values for the frequencies of the discoursal categories of MWCs (vertical lines indicate ± 95% CI)

level of individual MWCs, and their frequencies were normalised per 500 words per text in the L2 writers' essays. In order to answer the second research question, the model for the structural categories of MWCs included all the MWCs, including those that did not occur in the L2 writers' essays and those that occurred in the academic reading input (this also applies to the discoursal categories of MWCs in the next section). Rather than pairwise differences of the structural categories of MWCs, this section focuses on the effects of normalised frequency and dispersion of MWCs in the input corpus on the normalised frequency of MWCs in the L2 writers' essays.

There was a significant effect of the frequency of MWCs in the input corpus on the frequencies of NP-based MWCs in the L2 writers' essays, suggesting that NP-based MWCs that frequently occurred in the reading materials of L2 writers, were

Table 9 Descriptive statistics for the normalised frequencies of structural categories of MWCs and DPnorm

		Month 3	Month 5	Month 9
NP-based MWCs $N = 92$	Frequency in essays	0.02 (0.03)*	0.03 (0.04)	0.03 (0.04)
	Frequency in input corpus	0.03 (0.03)	0.03 (0.03)	0.03 (0.03)
	DPnorm	0.72 (0.16)	0.72 (0.16)	0.74 (0.14)
PP-based MWCs $N = 40$	Frequency in essays	0.04 (0.11)	0.03 (0.09)	0.03 (0.09)
	Frequency in input corpus	0.02 (0.03)	0.02 (0.03)	0.02 (0.03)
	DPnorm	0.81 (0.19)	0.78 (0.19)	0.79 (0.17)
VP-based MWCs $N = 97$	Frequency in essays	0.01 (0.03)	0.02 (0.03)	0.02 (0.02)
	Frequency in input corpus	0.01 (0.01)	0.01 (0.01)	0.02 (0.02)
	DPnorm	0.80 (0.16)	0.77 (0.16)	0.83 (0.14)

*Mean (Standard deviation)

used more frequently by the L2 writers, irrespective of time, as seen in Table 10. Similarly, the L2 writers used PP-based MWCs that they were frequently exposed to in their reading materials more in their essays, as there was a significant interaction between PP-based MWCs and the frequency of MWCs in the input corpus (see Fig. 3). Indeed, the frequency effects were strongest for PP-based MWCs, as shown in Table 10. No significant interaction was observed between the frequency of MWCs in the input corpus and VP-based MWCs, and pairwise comparisons indicated that the frequency of VP-based MWCs in the input corpus had no significant effect on their frequency in the L2 writers' essays ($t = -1.97$, $p = 0.28$). These frequency effects were observed irrespective of time and DPnorm. A significant interaction between DPnorm and 'Month 9' was observed in the frequency of MWCs in the L2 writers' essays, showing that MWCs that dispersed more evenly in all the compulsory reading resources occurred more frequently in the L2 writers' essays at Month 9. This effect was observed irrespective of the frequency of MWCs in the input corpus and the structural categories of MWCs. Non-significant effect of DPnorm on 'Month 3' and non-significant interaction between DPnorm and 'Month 5' indicated that DPnorm had no effect on the frequency of MWCs in the L2 writers' essays at Month 3 or at Month 5. It should be noted that the variable of the writers' L2 proficiency was dropped from the final model, along with three- and four-way interactions between the variables during the model selection, since the model

Table 10 Mixed-effects model for the structural categories of individual MWCs in the learner corpus

Parameters	Fixed effects			
	Estimate	SE	t	p
(Intercept)*	0.03	0.01	2.77	0.006
Frequency input	0.36	0.08	4.45	<0.001
PP-based MWCs	−0.02	0.007	−3.19	0.002
VP-based MWCs	0.0009	0.005	0.17	0.86
DPnorm	−0.019	0.01	−1.59	0.11
Month 5	−0.0002	0.007	−0.03	0.98
Month 9	0.02	0.008	2.65	0.008
Frequency input: PP-based MWCs	1.63	0.15	10.99	<0.001
Frequency input: VP-based MWCs	−0.19	0.11	−1.70	0.09
DPnorm: Month 5	−0.0005	0.009	−0.05	0.96
DPnorm: Month 9	−0.03	0.01	−2.53	0.01
	Random effects			
	Variance		SD	Corr
(Intercept)|MWCs	0.0009		0.03	
(Intercept)|ID	0.00002		0.005	
PP-based MWCs|ID	0.0001		0.011	−0.12
VP-based MWCs|ID	0.00001		0.004	−0.77

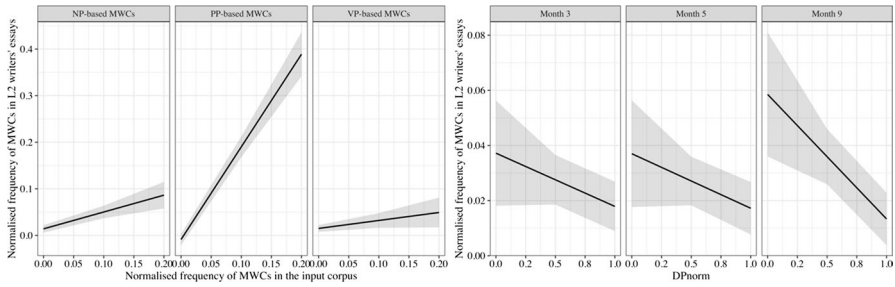*Reference level (intercept) is NP-based MWCs at Month 3

Fig. 3 Predicted values for the normalised frequencies of the structural categories of individual MWCs in L2 writers' essays across the frequencies of MWCs in the input corpus, time and DPnorm values (shaded areas indicate $\pm 95\%$ CI)

complexity was reduced until a further reduction indicated a bigger AIC value (see Murakami, 2016).

The random effects in Table 10 show that the frequency of MWCs showed variance across the different MWCs, and a parametric bootstrap test indicated significant random intercepts for the MWCs ($p < 0.001$). For the L2 writers, the random slopes were statistically significant ($p = 0.01$), but the random intercepts were not significant ($p = 0.23$). R^2 marginal value of the model was 0.46, and R^2 conditional value of the model was 0.50, suggesting that 50% of the variance in the frequency of structural categories of MWCs was explained by the model.

Discoursal categories of MWCs. Descriptive statistics for the normalised frequency and DPnorm of discoursal categories of MWCs can be seen in Table 11. It should be noted that the normalised frequency figures of MWCs are very low because the analysis was conducted at the level of individual MWCs, and their frequencies were normalised per 500 words per text in the L2 writers' essays. This section focuses on the effects of normalised frequency and dispersion of MWCs in the input corpus on the normalised frequency of discoursal categories of MWCs in the

Table 11 Descriptive statistics for the normalised frequencies of discoursal categories of MWCs and DPnorm

		Month 3	Month 5	Month 9
Referential expressions $N = 126$	Frequency in essays	0.03 (0.05)	0.03 (0.04)	0.03 (0.04)
	Frequency in input corpus	0.02 (0.03)	0.03 (0.02)	0.03 (0.03)
	DPnorm	0.74 (0.17)	0.73 (0.16)	0.75 (0.15)
Discourse organisers $N = 35$	Frequency in essays	0.04 (0.10)	0.03 (0.10)	0.03 (0.09)
	Frequency in input corpus	0.02 (0.04)	0.02 (0.04)	0.03 (0.05)
	DPnorm	0.80 (0.19)	0.78 (0.18)	0.80 (0.17)
Stance expressions $N = 76$	Frequency in essays	0.01 (0.02)	0.01 (0.03)	0.02 (0.02)
	Frequency in input corpus	0.01 (0.02)	0.02 (0.02)	0.02 (0.02)
	DPnorm	0.80 (0.16)	0.77 (0.16)	0.83 (0.14)

* Mean (Standard deviation)

L2 writers' essays rather than pairwise frequency differences between the different discoursal categories of MWCs.

The mixed-effects model results revealed that there was a significant effect of the frequency of MWCs in the input corpus on the frequency of referential expressions in the L2 writers' essays, suggesting that L2 writers used referential expressions that occurred frequently in their reading materials more, irrespective of time. Similarly, a significant interaction between the frequency of MWCs in the input corpus and discourse organisers showed that discourse organisers that occurred more frequently in the input corpus were used more by the L2 writers in their essays. The frequency effects were most pronounced for discourse organisers, as Fig. 4 shows. There was a non-significant interaction between the frequency of MWCs in the input corpus and stance expressions, and the pairwise comparisons showed that the frequency of stance expressions in the input corpus had no effect on their frequency in the L2 writers' essays ($t = -0.23$, $p = 0.99$). These frequency effects were present irrespective of time and DPnorm values. As shown in Table 12, DPnorm had a significant effect on the frequency of MWCs in the L2 writers' essays at Month 3, indicating that L2 writers used MWCs that were more sparsely dispersed in the input corpus less frequently in their essays. There was a non-significant interaction between DPnorm and Month 5, and the pairwise comparisons indicated the effect of DPnorm was significant at Month 5 ($p = 0.02$). Strikingly, a significant interaction between DPnorm and 'Month 9' showed that the effects of DPnorm were strongest at Month 9, as seen in Fig. 4. The effects of DPnorm were observed irrespective of the frequency of MWCs in the input corpus and their discoursal categories. As explained earlier, the variable of the L2 writers' proficiency was dropped from the final model, along with three- and four-way interactions between the variables during the model selection, since the model complexity was reduced until a further reduction indicated a bigger AIC value.

As Table 12 shows, there was variance explained by the individual MWCs in the frequency of discoursal categories of MWCs, and a parametric bootstrap test revealed that the random intercepts for the MWCs were statistically significant ($p < 0.001$). For the L2 writers, the random slopes were statistically significant ($p = 0.02$), and the random intercepts were non-significant ($p = 0.32$). R^2 marginal
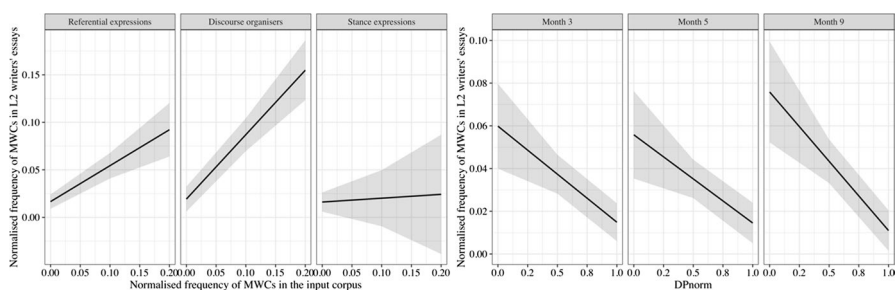

Fig. 4 Predicted values for the normalised frequencies of the discoursal categories of the individual MWCs in L2 writers' essays across the frequencies of MWCs in the input corpus, time and DPnorm values (shaded areas indicate ± 95% CI)

Table 12 Mixed-effects model for the discoursal categories of individual MWCs in the learner corpus

Parameters	Fixed effects			
	Estimate	SE	t	p
(Intercept)*	0.05	0.01	4.69	<0.001
Frequency_input	0.38	0.08	4.76	<0.001
Discourse organisers	0.003	0.008	0.35	0.72
Stance expressions	-0.0004	0.006	−0.07	0.95
DPnorm	−0.05	0.01	−3.56	<0.001
Month 5	−0.004	0.007	−0.57	0.57
Month 9	0.02	0.008	2.00	0.04
Frequency_input: Discourse organisers	0.3	0.11	2.71	0.007
Frequency_input: Stance expressions	−0.34	0.18	−1.89	0.06
DPnorm: Month 5	0.004	0.009	0.40	0.69
DPnorm: Month 9	−0.02	0.01	−1.96	0.04
	Random effects			
	Variance	SD		Corr
(Intercept)|MWCs	0.001	0.03		
(Intercept)|ID	0.00001	0.004		
Discourse organisers|ID	0.0001	0.01		−0.25
Stance expressions|ID	0.00001	0.003		−0.84

*Reference level (intercept) is Referential expressions at Month 3

value was 0.44, and R^2 conditional value was 0.51. This means that 51% variance in the frequency of discoursal categories of MWCs was explained by the whole model.

Discussion

In reference to the first research question on the changes in the frequency of MWCs, the present study showed that different categories of MWCs underwent distinct patterns of change over time. There was a significant increase in the frequency of NP-based MWCs in the L2 writers' essays from Month 3 to Month 9. Although there was a decreasing trend of PP-based MWCs and an increasing trend of VP-based MWCs, these trends were not statistically significant. It is worth noting that the effect size for the LMM for the structural categories of MWCs was smaller than the other models, suggesting that there were probably other factors that could account for variation in the frequency of structural categories of MWCs. It was also observed that the L2 writers used referential expressions more frequently at Month 9 than at Month 3. On the other hand, the L2 writers used discourse organisers less frequently at Month 9 in comparison to Month 3. This finding may not be surprising, since Appel and Wood (2016) found that L2 writers with a high level of writing proficiency used discourse organisers less frequently than those with a low level of writing proficiency, which could be interpreted as a developmental pattern for

writing at Month 9 in this study, though an in-depth qualitative analysis is necessary to examine qualitative changes. Similar to the patterns of change in VP-based MWCs (mostly stance expressions), an increasing trend was observed for stance expressions, but this change was not statistically significant. These findings reinforce the view that the use of MWCs undergoes slow-paced patterns of change in longitudinal L2 corpora (see Paquot & Granger, 2012; Siyanova-Chanturia & Spina, 2020), since the significant changes took place from Month 3 to Month 9 rather than from Month 3 to Month 5 or Month 5 to Month 9 for the frequencies of NP-based MWCs, referential expressions, and discourse organisers in the L2 writers' essays. It should be noted that there was an unequal time interval between the data collection time points. It remains unknown whether the significant changes in the frequencies of MWCs would still be found if there had been an equal time interval between the data collection time points and the essays had been collected at Month 7 instead of at Month 9.

The significant increase in NP-based MWCs and referential expressions in the L2 writers' essays at Month 9 suggests that the distribution of different categories of MWCs in the L2 writers' essays became closer to the distributional characteristics of MWCs in academic prose of English, since Biber et al. (2004) found that academic writing in English relies on noun phrases and prepositional phrases, most of which serve as referential expressions. This shows that that there was an association between referential expressions and NP-based MWCs in this study, which accords with Biber et al.'s (2004) findings. The evidence for increase in NP-based MWCs and referential expressions in the L2 writers' essays also gives support for usage-based approaches to language learning (e.g., Crossley et al., 2019; Ellis, 2002; Ellis et al., 2016) in that the L2 writers of this study became more attuned to the distributional characteristics of MWCs in academic writing at the end of the academic year. This seemed to be shaped by their cumulative encounters of MWCs in their reading and writing, since they were asked to read 74 texts in their compulsory course units and submit their assignments in English during their first year at university. The decreasing trend of PP-based MWCs could be traced back to a significant decrease in the frequency of discourse organisers in the L2 writers' essays, since many of the PP-based MWCs, such as 'in addition to' were discourse organisers that decreased over time in this study. This suggests an association between the frequencies of discourse organisers and PP-based MWCs in the present study.

The random effects for models for both the structural and discoursal categories of MWCs revealed that there was variation across the L2 writers in the use of MWCs at the levels of both initial status (intercept) and rate of change (slope), and the random slopes were statistically significant. Additionally, negatively correlated random effects were observed in that the essays that contained one category of MWCs less frequently included more of these MWCs and vice versa over time. For example, the L2 writers that used VP-based MWCs less frequently at Month 3 used them increasingly more frequently over time. This highlights the heterogeneity of the L2 writers of this study in terms of the rate of change in the frequency of MWCs, though the L2 writers constituted a single cohort in the same programme from very similar backgrounds, taking the same compulsory course units. The considerable individual variation found in this study underscores the importance of using mixed-effects

modelling to account for sources of random variation (e.g., Gries, 2015; Murakami, 2016; Siyanova-Chanturia & Spina, 2020). The individual variation found in this study is not in line with the findings of Siyanova-Chanturia and Spina (2020) who found very little individual variation in the use of noun + adjective combinations in the essays of 175 L1 Chinese learners of Italian, and the individual variation was only at the initial status. These different findings may be traced back to the different participant samples and their characteristics. For example, the L2 writers had advanced proficiency of English (B2/C1 CEFR levels) in this study, whereas L1 Chinese learners of Italian had beginner-, elementary-, and intermediate-level proficiency of Italian in Siyanova-Chanturia and Spina's study (2020). All the L2 writers in this study were advanced learners of English; therefore, the findings of this study may not be generalised to L2 writers with other proficiency levels. This study found significant changes in the frequencies of NP-based MWCs, referential expressions and discourse organisers only from Month 3 to Month 9, which suggests a rather slow pattern of change in MWCs in advanced L2 writing. There is evidence that "beginner learner collocational knowledge can improve over a relatively short period of time" (Siyanova-Chanturia, 2015, p. 158). Indeed, Siyanova-Chanturia (2015) reported longitudinal improvement in noun-adjective combinations in beginner-level L2 writing of L1 Chinese learners of Italian over five months, though it is not possible to compare these studies due to the different characteristics, including the participants' L1 background, research context, and multi-word constructions that were investigated.

With regard to the second research question, this study revealed that there were significant frequency effects for the frequencies of all categories of MWCs, except for VP-based MWCs and stance expressions in the L2 writers' essays. These frequency effects of MWCs in the input corpus on the frequency of MWCs in the L2 writers' essays were observed irrespective of time. This finding is not in line with those that have been reported in previous longitudinal studies in L2 writing (Bestgen & Granger, 2014; Siyanova-Chanturia & Spina, 2020). Using different methodology from this study, these previous studies showed that L2 writers used fewer frequent collocations over time. In spoken L2 studies, an opposite trend was observed, since it was reported that L2 learners increasingly relied on high-frequency *n*-grams (Garner & Crossley, 2018; Kim et al., 2018). These different findings could be attributed to the methodological differences and/or mode of discourse, since spoken discourse relies on MWCs much more than written discourse, as Biber et al. (2004) found. The findings of frequency effects for this study are not consistent with those of Siyanova-Chanturia (2015) who found that beginner level L2 writers employed more higher frequency collocations over the period of four months. It should be noted that these frequency effects of MWCs in this study are not directly comparable with the above-mentioned studies due to the different L1 backgrounds of the participants, research environments, proficiency levels and methods used; therefore, the comparisons of the frequency effects should be treated with caution. In this study, there was also a great deal of inter-MWC variation in the frequency of MWCs in the L2 writers' essays. In addition to frequency, dispersion was found to significantly predict the frequency of MWCs in the L2 writers' essays at Month 9 for the structural categories of MWCs and across all the time points (strongest at Month 9) for the discoursal

categories of MWCs, irrespective of the frequency of MWCs in the input corpus. This suggests that when MWCs dispersed more evenly in the input corpus, the L2 writers used them more frequently in their essays, especially at Month 9. Hence, this study gives empirical evidence for the importance of dispersion in learning to use MWCs in L2 writing longitudinally. The findings reinforce the view that frequency alone cannot be the only measure for exposure, and dispersion should also be considered in addition to frequency information in learner corpus studies (Gablasova et al., 2017; Gries, 2010).

Taken together, the two models that were built to address the second research question included the structural/discoursal categories of MWCs, the frequency and dispersion of MWCs in the academic reading input as well as time and explained greater variance in the frequency of MWCs in the L2 writers' essays than the first two models that included only time and structural/discoursal categories of MWCs. This shows that the frequency of MWCs in the L2 writers' essays became tuned by both time and L2 writers' 'usage' experiences, as operationalised by the frequency and dispersion of MWCs in the input corpus, which is in line with the usage-based approaches to language learning (Ellis et al., 2016; Ellis & Wulff, 2015). It is striking that time interacted with dispersion in that dispersion effects were only statistically significant at Month 9 for the structural categories of MWCs and that these effects were strongest at Month 9 for the discoursal categories of MWCs. This suggests that the connection between the MWCs in the L2 writers' input and the MWCs in the L2 writers' essays became stronger over a long period of time, through repeated and regular encounters with the MWCs dispersed in a range of reading materials.

The frequencies of MWCs tended to be rare in the L2 essay writers' essays, though this should be treated with caution since no reference corpus was used in this study. Furthermore, this study only examined target-like academic MWCs rather than all possible MWCs. Overall, the findings are in line with usage-based approaches to language learning in that the more frequently and regularly the L2 learners were exposed to a construction in their reading resources over time, the more frequently they used them in their academic writing, though frequency effects could vary across the categories of MWCs, and dispersion effects could vary across time. This supports the view that "reading and writing are reciprocal activities" (Grabe & Kaplan, 1996, p. 297), and academic reading served as one source of input for MWCs used in the L2 writers' essays.

Limitations and future work

This study has several limitations that should be addressed in future studies. First, the results were based on the quantitative analysis of MWCs in the L2 writers' essays; therefore, no claims can be made regarding their appropriacy in context. Second, it was not possible to include the grades of the L2 writers' assignments in modelling the frequencies of MWCs. A further study could research the relationship between the writing quality/grades and frequency of MWCs over time. The present study provided empirical evidence for the effects of time, frequency and dispersion

of the MWCs in the reading resources on the frequency of MWCs in the L2 writers' essays over time. However, due to the non-experimental nature of this study, there may have been other learner-related uncontrollable factors, such as the use of reference sources at the time of writing, other kinds of input the learners were exposed to, and out-of-class exposure to English as well as other variables related to MWCs, including salience and concreteness (see Ellis, 2012) that could have accounted for variance in the frequency of MWCs in the L2 writers' essays. In this study, ecological validity was prioritised, and it was not possible to control these abovementioned factors. It should be acknowledged that "there is often a trade-off between the degree to which research studies reflect the realities of a research context [ecological validity] versus the degree of control" (Loewen & Plonsky, 2016, p. 56) over extraneous variables. Therefore, future studies should take these other variables and their interactions into account. For example, further research might trace L2 writers' input that they receive during lectures and their out-of-class exposure to English and test their proficiency of English at each data collection point in order to better explain individual trajectories, since random slopes for the L2 writers were statistically significant in this study. Finally, a small longitudinal learner corpus with three waves of data was used in the study, and there was an unequal time interval between the data collection time points (Month 3, Month 5, and Month 9). Further research using a larger longitudinal corpus of L2 academic writing, with denser waves of data (see Verspoor, Lowie, & de Bot, 2011) at regular and equal time intervals over more than one academic year is needed to track the trajectories of MWCs.

Conclusion

The present study revealed changes in the frequency of MWCs across different functional and structural categories of MWCs, by using linear mixed-effects modelling. The significant increases in NP-based MWCs and referential expressions suggest that the L2 writers at Month 9 overall approximated to the use of MWCs in English academic prose in quantitative and distributional terms, since Biber et al. (2004) note that referential expressions are the most commonly occurring discoursal category of word sequences in English academic writing and that referential expressions are mostly comprised of noun and prepositional phrases. It is likely that the L1 Turkish learners of English became sensitive to the distributional characteristics of MWCs in academic writing over time (Ellis, 2002), due to their increasing experience of reading and writing academic sources and their exposure to English at an English-medium university, which contributed to their incremental learning of MWCs. This is the first study that provided empirical evidence for significant effects of frequency and dispersion of MWCs in the reading resources of the L2 writers on the frequency of MWCs in the L2 writers' essays, by tracing their input as part of their curriculum. These suggest that frequency and dispersion both play an important role in learning to use MWCs in L2 writing. These findings seem to be encouraging given that the L2 writers of this study arguably had more limited exposure to English than their counterparts in immersion settings.

The new findings of the present study have important implications for material design and teaching MWCs. Given that both frequency and dispersion of MWCs in the input corpus predicted the frequency of MWCs, L2 teaching materials should be designed in a way that would give students ample opportunities for exposure to target constructions frequently and regularly in a wide range of contexts. The significant dispersion effects of the structural categories of MWCs at Month 9 and discoursal categories of MWCs across the data collection points (strongest at Month 9) in the academic reading input on the frequency of MWCs in the L2 writers' essays suggest wider educational implications for L2 reading and teaching MWCs. In L2 reading resources, such as graded readers, MWCs should be dispersed widely so that learners can receive repeated and regular exposure to them across different contexts over time. This implication may also be extended to classroom settings in that teachers could cover and repeat MWCs over a number of different sessions rather than teaching them in one session. Furthermore, as this study revealed a slow developmental pattern for the frequency of NP-based MWCs and referential expressions, which are the building blocks of academic prose in English, their explicit instruction may be necessary for L2 writers to use them in their writing at English-medium universities. The frequency of MWCs in the input of academic reading had no significant effect on the frequency of stance expressions or VP-based MWCs in the L2 writers' essays; therefore, these can be explicitly taught through the use of corpus-based activities in class. For instance, freely available web-based academic corpora can be used to study the most frequently used MWCs and to explore their discourse functions in class.

References

Ambridge, B., Theakston, A., Lieven, E., & Tomasello, M. (2006). The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development, 21,* 174–193. https://doi.org/10.1016/j.cogdev.2005.09.003.

Appel, R., & Wood, D. (2016). Recurrent word combinations in eap test-taker writing: differences between high- and low-proficiency levels. *Language Assessment Quarterly, 13,* 55–71. https://doi.org/10.1080/15434303.2015.1126718.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Bartoń, K. (2019). *MuMIn: Multi-Model Inference*. R package version 1.43.15. https://CRAN.R-project.org/package=MuMIn

Bates, D. M., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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. https://doi.org/10.1016/j.jslw.2014.09.004.

Bi, P. (2020). Revisiting genre effects on linguistic features of L2 writing: A usage-based perspective. *International Journal of Applied Linguistics*, 1–16. https://doi.org/10.1111/ijal.12297

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

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25,* 371–405. https://doi.org/10.1093/applin/25.3.371.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics, 20,* 139–173. https://doi.org/10.1075/ijcl.20.2.01bre.

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers TESOL, 5,* 31–64.

Candarli, D. (2020). Changes in L2 writers' self-reported metalinguistic knowledge of lexical phrases over one academic year. *The Language Learning Journal*, *48*, 768–784. https://doi.org/10.1080/09571736.2018.1520914.

Chen, Y. H., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics, 37,* 849–880. https://doi.org/10.1093/applin/amu065.

Coffin, C. (1996). *Exploring literacy in school history*. Sydney: NSW Department of School Education.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23,* 397–423. https://doi.org/10.1016/j.esp.2003.12.001.

Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes, 12,* 33–43. https://doi.org/10.1016/j.jeap.2012.11.002.

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,* 1–26. https://doi.org/10.1515/iral.2011.001.

Crossley, S., Skalicky, S., Kyle, K., & Monteiro, K. (2019). Absolute frequency effects in second language lexical acquisition. *Studies in Second Language Acquisition, 41,* 721–744. https://doi.org/10.1017/S0272263118000268.

Crossley, S., Subtirelu, N., & Salsbury, T. (2013). Frequency effects or context effects in second language word learning. *Studies in Second Language Acquisition, 35,* 727–755. https://doi.org/10.1017/S0272263113000375.

Durrant, P. (2017). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics, 38,* 165–193. https://doi.org/10.1093/applin/amv011.

Durrant, P., & Brenchley, M. (2019). Development of vocabulary sophistication across genres in English children's writing. *Reading and Writing: An Interdisciplinary Journal, 32,* 1927–1953. https://doi.org/10.1007/s11145-018-9932-8.

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. https://doi.org/10.1515/iral.2009.007.

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. https://doi.org/10.1017/S0272263102002024.

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

Ellis, N. C., & Ferreira-Junior, F. (2009). Construction learning as a function of frequency, frequency distribution, and function. *Modern Language Journal, 93,* 370–386. https://doi.org/10.1111/j.1540-4781.2009.00896.x.

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar*. Language Learning Monograph Series. Malden, MA: Wiley.

Ellis, N.C., & Wulff, S. (2015). Usage-based approaches to SLA. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 75–93). New York, NY: Routledge.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning, 67*(S1), 130–154. https://doi.org/10.1111/lang.12226.

Garner, J., & Crossley, S. (2018). A latent curve model approach to studying L2 n-gram development. *Modern Language Journal, 102,* 494–511. https://doi.org/10.1111/modl.12494.

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

Grabe, W. & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective*. London: Longman.

Granger, S., Gilquin, G., & Meunier, F. (2016). *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press.

Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics, 13,* 403–437. https://doi.org/10.1075/ijcl.13.4.02gri.

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

Gries, S. Th. (2015). The most under-used statistical method in corpus linguistics: Multilevel (and mixed-effects) models. *Corpora, 10,* 95–125. https://doi.org/10.3366/cor.2015.0068.

Halekoh, U., & Højsgaard, S. (2014). A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest. *Journal of Statistical Software*, *59*(9), 1–30. https://doi.org/10.18637/jss.v059.i09

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27,* 4–21. https://doi.org/10.1016/j.esp.2007.06.001.

Hyland, K. (2013). Writing in the university: education, knowledge and reputation. *Language Teaching, 46,* 53–70. https://doi.org/10.1017/S0261444811000036.

Kim, M., Crossley, S., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *Modern Language Journal, 102,* 120–141. https://doi.org/10.1111/modl.12447.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34,* 12–24. https://doi.org/10.1016/j.jslw.2016.10.003.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Larsen-Freeman. D. (2015). Complexity theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 227–244). New York, NY: Routledge.

Leedham, M., & Cai, G. (2013). Besides … on the other hand: Using a corpus approach to explore the influence of teaching materials on Chinese students' use of linking adverbials. *Journal of Second Language Writing, 22,* 374–389. https://doi.org/10.1016/j.jslw.2013.07.002.

Lenth, R. V. (2016). Least-squares means: the R package lsmeans. *Journal of Statistical Software, 69*(1), 1–33. org/https://doi.org/10.18637/jss.v069.i01

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing, 18,* 85–102. https://doi.org/10.1016/j.jslw.2009.02.001.

Lijffijt, J., & Gries, S. Th. (2012). Correction to Stefan Th. Gries' "Dispersions and adjusted frequencies in corpora", International Journal of Corpus Linguistics. *International Journal of Corpus Linguistics, 17,* 147–149. https://doi.org/10.1075/ijcl.17.1.08lij.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning, 65*(S1), 185–207. https://doi.org/10.1111/lang.12117.

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes, 31,* 25–35. https://doi.org/10.1016/j.esp.2011.07.002.

Loewen, S., & Plonsky, L. (2016). *An A-Z of applied linguistics research methods*. New York, NY: Palgrave.

Lüdecke, D., Makowski, D., & Waggoner, P. (2019). *performance: assessment of regression models performance. R package version 0.4.0*. https://CRAN.R-project.org/package=performance

Murakami, A. (2016). Modeling systematicity and individuality in nonlinear second language development: The case of English grammatical morphemes. *Language Learning, 66,* 834–871. https://doi.org/10.1111/lang.12166.

Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge, England: Cambridge University Press.

Ooms, J. (2018). *tesseract: Open Source OCR Engine*. R package version 4.0. https://CRAN.R-project.org/package=tesseract

Ortega, L. (2018). Ontologies of language, second language acquisition, and world Englishes. *World Englishes, 37,* 64–79. https://doi.org/10.1111/weng.12303.

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics, 25,* 26–45. https://doi.org/10.1017/S0267190505000024.

Paquot, M. (2018). Phraseological competence: A missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly, 15,* 29–43. https://doi.org/10.1080/15434303.2017.1405421.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32,* 130–149. https://doi.org/10.1017/S0267190512000098.

R Core Team (2019). *R: A language and environment for statistical computing* [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research, 8*(1), 3–15.

Scott, M. (2012). *WordSmith tools (Version 6.0)*. Liverpool: Lexical Analysis Software.

Selvi, A. F. (2020). Resisting English medium instruction through digital grassroots activism. *Journal of Multilingual and Multicultural Development, 1–17,*. https://doi.org/10.1080/01434632.2020.1724120.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31,* 487–512. https://doi.org/10.1093/applin/amp058.

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

Siyanova-Chanturia, A., & Spina, S. (2020). Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study. *Language Learning, 70,* 420–463. https://doi.org/10.1111/lang.12383.

Taylor, L. (2004). IELTS, Cambridge ESOL examinations and the common European framework. *Research Notes, 18,* 2–3.

Verspoor, M., Lowie, W., & de Bot, K. (Ed.). (2011). *A dynamic approach to second language development: Methods and techniques*. Amsterdam, Netherlands: John Benjamins.

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

Wulff, S. (2019, September). *Getting more use out of "usage"*. Paper presented at Second Language Research Forum (SLRF), Michigan State University, US. Retrieved from http://slrf2019.sls.msu.edu/dr-stefanie-wulff/

Yoon, H.-J. (2016). Association strength of verb-noun combinations in experienced NS and less experienced NNS writing: Longitudinal and cross-sectional findings. *Journal of Second Language Writing, 34,* 42–57. https://doi.org/10.1016/j.jslw.2016.11.001.

Zheng, Y. (2016). The complex, dynamic development of L2 lexical use: A longitudinal study on Chinese learners of English. *System, 56,* 40–53. https://doi.org/10.1016/j.system.2015.11.007.