LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Multi-Word Expressions in Second Language Writing: A Large-Scale Longitudinal Learner Corpus Study

Anna Siyanova-Chanturia [a] and Stefania Spina [b]

[a]Victoria University of Wellington and [b]University for Foreigners of Perugia

In the present study, we sought to advance the field of learner corpus research by tracking the development of phrasal vocabulary in essays produced at two different points in time. To this aim, we employed a large pool of second language (L2) learners ($N = 175$) from three proficiency levels—beginner, elementary, and intermediate—and focused on an underrepresented L2 (Italian). Employing mixed-effects models, a flexible and powerful tool for corpus data analysis, we analyzed learner combinations in terms of five different measures: phrase frequency, mutual information, lexical gravity, delta $P_{forward}$, and delta $P_{backward}$. Our findings suggest a complex picture, in which higher proficiency and greater exposure to the L2 do not result in more idiomatic and targetlike output, and may, in fact, result in greater reliance on low frequency combinations whose constituent words are non-associated or mutually attracted.

Keywords multi-word expressions; learner corpora; longitudinal; Italian; second language acquisition; mixed-effects modeling

Introduction

Interest in learner corpus work has surged in recent years (Granger, 2019; Granger, Gilquin, & Meunier, 2015; Paquot & Granger, 2012). Much of this research has, in particular, focused on the use and development of second language (L2) vocabulary, both single words and longer stretches of language.

The research leading to these findings was supported by Victoria University of Wellington grants no. 210037 and no. 213857 to Anna Siyanova-Chanturia and by the University for Foreigners of Perugia with a 2018 grant to Stefania Spina for the research project *Differenti tipologie di combinazioni lessicali e acquisizione dell'italiano come L2*.

Correspondence concerning this article should be addressed to Anna Siyanova-Chanturia, School of Linguistics and Applied Language Studies, Victoria University of Wellington, Kelburn Parade, Wellington, 6012, New Zealand. E-mail: anna.siyanova@vuw.ac.nz

© 2019 Language Learning Research Club, University of Michigan
DOI: 10.1111/lang.12383

However, much of this work has been done with learner data collected at one point in time. Longitudinal studies are still uncommon, and researchers have called for a greater emphasis on studies conducted over a period of time with the same group of learners (Bestgen & Granger, 2014; Laufer & Waldman, 2011; Paquot & Granger, 2012).

In the present study, we sought to advance the understanding of the development of a L2, by tracing the trajectory of multi-word expression use over a period of time. Multi-word expressions are sequences above the word level that may vary along the continua of frequency, length, fixedness, abstractness, and figurativeness/literality, and a proficient language user may recognize them as conventional (Siyanova-Chanturia & van Lancker Sidtis, 2019). Multi-word expressions such as collocations, idioms, binomials, lexical bundles, and other phrasal elements are ubiquitous in language (Jackendoff, 1995; Langacker, 1987; Tomasello, 2003). Critically, they are an integral part of the mental lexicon on a par with single words and have thus been claimed to be fundamental building blocks of language (Arnon, McCauley, & Christiansen, 2017; Arnon & Snider, 2010; Christiansen & Chater, 1999; Elman, 2009; Wray, 2002). Yet, as a great many studies can attest, L2 learners often experience considerable difficulties acquiring and using multi-word expressions in speech and writing (Foster, Bolibaugh, & Kotula, 2014; Wray, 2002).

Researchers have proposed that first language (L1) learners and adult speakers differ from L2 learners in their reliance on multi-word information. Although L1 speech is chunked in nature, L2 speech is markedly less so. In a recent study employing a computational model on corpus data from L1 learners, L1 adult speakers, and L2 adult speakers, McCauley and Christiansen (2017) found that the L2 adults' speech was characterized by a lesser use of multi-word expressions than was the speech of the other groups. This finding is in accord with Wray's (2002, 2019) and Arnon and Christiansen's (2017) propositions that multi-word expressions play a different role in L1 and L2 learning, processing, and use, with nonnative speakers generally relying less on multi-word information. One of the reasons for such asymmetry might be the fact that in L2 learning, the focus has traditionally been on aiding learners to amass single words and grammatical rules to enable them to produce a seemingly infinite number of novel utterances. In addition, L2 learners are rarely made aware of combinatorial mechanisms in language, that is, the fact that any one word may only have a rather limited number of words (collocates) which it can be juxtaposed with. Of course, neither are L1 learners made explicitly aware of such constraints. However, unlike L2 learners, L1 learners have had huge amounts of exposure to the myriads of phrasal configurations, which allows them to

know that, for example, *heavy rain* and *strong wind* are more idiomatic and nativelike than the seemingly identical *strong rain* and *heavy wind.* Given their often rather limited exposure to and experience with a language, L2 learners are often said to rely on linguistic creativity and make "overliberal assumptions about the collocational equivalence of semantically similar items" (Wray, 2002, pp. 201–202).

The dissimilarities in the treatment of multi-word information by L1 and L2 speakers, as McCauley and Christiansen (2017), Wray (2002, 2019) and other researchers have proposed, have been reported in numerous learner corpus studies. By far the most commonly reported finding has been underuse or overuse of multi-word expressions by L2 learners compared to a L1 baseline (Altenberg & Granger, 2001; Chen & Baker, 2010; De Cock, 2004; Durrant & Schmitt, 2009; Gilquin, 2007; Granger, 1998; Granger & Paquot, 2009; Henderson & Barr, 2010; Howarth, 1998; Laufer & Waldman, 2011; Nessalhauf, 2005). And even where researchers found the quantity of target L2 structures to be on a par with L1 speakers, they found that the quality, or native-likeness, of L2 items fell short of L1 norms (Siyanova-Chanturia & Schmitt, 2008). The studies in the field have also attested to a negative influence of learners' L1 on L2 phraseological performance, in particular where L1 and L2 belong to different language families (Altenberg & Granger, 2001; Gilquin, 2007; Granger & Paquot, 2009; Henderson & Barr, 2010; Lorenz, 1999), and they have reported an important role for immersion-based exposure to the target structures, with the time spent in a L2 country being a significant predictor of the quality and quantity of L2 multi-word expressions (Groom, 2009; Siyanova-Chanturia & Schmitt, 2008; Waibel, 2008). At the same time, researchers have also noted that higher proficiency and greater experience with the L2 does not necessarily result in more accurate use of target multi-word expressions (Laufer & Waldman, 2011; Nesselhauf, 2005; but see Crossley & Salsbury, 2011; Li, Eskildsen, & Cadierno, 2014; Siyanova-Chanturia, 2015; Yuldashev, Fernandez, & Thorne, 2013).

In the present study, we sought to shed further light on the complexities involved in the acquisition and use of multi-word information in a L2. Capitalizing on the currently available body of knowledge and the major gaps in the field of longitudinal learner corpus research, we collected and analyzed written corpus data from a large participant pool, focusing, in particular, on N+Adj word combinations produced by learners of L2 Italian from three different proficiency levels. We first review available learner corpus research—case and larger-scale studies—in which researchers have explored the development of multi-word expressions over a period of time.[1] We then briefly turn to the

statistical methods that are best suited for the analysis of longitudinal corpus data—a question that has become increasingly important and to which recent calls for more rigorous approaches to data analysis guided our analyses. Finally, we outline the rationale for the current study, focusing specifically on the gaps in the literature that have helped shape our approach before turning to the method, findings, and general discussion.

Background Literature

Longitudinal Studies Into the Acquisition and Use of Multi-Word Expressions in a Second Language

Case Studies

Because longitudinal studies are often logistically challenging, it is unsurprising that much of the research looking at the development of phrasal vocabulary in L2 learner writing has been based on case studies, employing one or only a handful of participants. In some of the earliest of such studies, Li and Schmitt (2009) and Li and Schmitt (2010) followed one and four Chinese English as a second language (ESL) learners, respectively, over a period of 12 months. Their focus was on the development of lexical phrases (e.g., Adj+N collocations) in the learners' master's essays and theses, which can be classified as formal academic discourse. Although Li and Schmitt (2009) observed some improvement in the quality and quantity of L2 phrasal usage in a single-learner study, Li and Schmitt's (2010) findings were less optimistic, attesting to little change in the learners' production of collocations and to a high degree of variability among the four learners.

In a more recent corpus study, Yuldashev et al. (2013) analyzed a different kind of written discourse—informal instant messages and blogs produced in out-of-class settings. In particular, the focus of this study was on the use of *es/que* ("[it] is/that") by three learners of L2 Spanish. Akin to Li and Schmitt (2010), they found some degree of variability among the learners (i.e., learners differed among themselves in the way that they used the target chunk). Unlike Li and Schmitt (2010), however, the three L2 learners showed significant improvement in their use of the chunk, being able to use it both in a fixed and increasingly schematic fashion.

In contrast to the researchers who explored phrasal production in written language in the above studies, Crossley and Salsbury (2011) and Li et al. (2014) investigated the development of phrasal vocabulary in a range of spoken texts, the former following six ESL learners over a period of 1 year, and the latter monitoring one ESL learner over the course of three and a half years. Tracing the learners' use of two-word lexical bundles (*going to, want to*) in casual

conversations, Crossley and Salsbury (2011) made two observations. First, the amount of time in an English-speaking country was an important predictor of how well the learners were able to use lexical bundles. Second, with time, the learners were able to use lexical bundles at frequencies that were comparable to those in L1 usage. This ability, in particular, was a significant finding, implying that 1 year might be sufficient for L2 learners to start exhibiting (at least some) nativelike speaking behaviors. Li et al. (2014) data further attested to improved phrasal usage over the course of time. Monitoring one learner in informal classroom interactions, Li et al. (2014) showed that the learner's use of L2 English motion constructions (*go to Mexico, come to the party*) became more productive, with new, emerging linguistic patterns actively building on previous experience (also see Myles, Hooper, & Mitchell, 1998).

The studies that we have reviewed are important in that, focusing on a small number of learners, they painted a detailed picture of L2 phrasal vocabulary development from a longitudinal perspective. Where researchers have observed more than one learner, they have also noted the role of individual variation, with learners exhibiting varying degrees of success in learning and using target structures over time. Interestingly, although some studies reported modest or no improvement (Li & Schmitt, 2009, 2010), others indicated significant gains both in written and spoken discourse (Li et al., 2014; Yuldashev et al., 2013). This is, perhaps, not surprising given that one of the limitations (and, admittedly, strengths) of case studies is that findings concern a rather limited number of language learners—and often just one learner—and are thus not generalizable to other L2 populations.

Large-Scale Studies

Unlike case studies, large-scale investigations are not able to capture every detail specific to the learning or use of a target structure at the level of an individual learner. However, such studies are more likely to provide a more accurate and, critically, generalizable and replicable picture of the mechanisms at play in vocabulary learning over time. Only a few studies to date have employed a (relatively) large L2 participant pool to investigate written production. We have reviewed these in some detail so as to provide comparisons with our large-scale longitudinal investigation.

The study by Qi and Ding (2011) was arguably the first one to use a sizable group of L2 learners. Combining longitudinal and cross-sectional approaches, these authors analyzed the use of formulaic expressions in monologues produced by 56 English as a foreign language (EFL) learners (L1 Chinese) at the beginning (Year 1) and at the end (Year 4) of a 3-year period. Qi and

Ding further compared these learners' performance with that of English native speakers. The researchers focused on a variety of word combinations (two or more words in length that operate as a phrase) and three aspects of their development over time: frequency, accuracy, and variation. Unlike the quantitative, frequency-based approach which researchers had adopted in other longitudinal studies (Bestgen & Granger, 2014; Siyanova-Chanturia, 2015; Yoon, 2016), in this study, Qi and Ding followed the phraseological approach for researching phrasal vocabulary. Proponents of the phraseological approached rely on dictionaries and native speaker judgments (i.e., their intuition) of formulaicity in the identification and extraction of collocation (Foster, 2001; Howarth, 1998; Lennon, 1990; Nesselhauf, 2003). Qi and Ding's (2011) results proved to be mixed. They observed no differences between Year 1 and Year 4 monologues for phrase frequency and accuracy. They also found that the L2 learners lagged behind the native speaker control group in frequency and accuracy. However, the learners showed a marked improvement in the variation in their oral production of expressions in Year 4 relative to Year 1, although many of these new expressions were used incorrectly. This was taken to indicate that, as EFL learners become more proficient and acquire a richer repository of phrasal expressions, they also err more frequently when using these newly acquired sequences.

In another study exploiting longitudinal and cross-sectional approaches, Yoon (2016) explored the development of V+N combinations (*find solution, get information*) in narrative and argumentative essays written by 51 ESL learners (from a variety of L1 backgrounds). The researcher adopted a frequency-based approach in which they used statistical indices of formulaicity—frequency and a range of association measures (also see Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Siyanova-Chanturia, 2015; Siyanova-Chanturia & Schmitt, 2008; Yoon, 2016). In the frequency-based tradition, collocation is often defined as "the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey, 1991, p. 7; also see Hoey, 2005; Jones & Sinclair, 1974; Manning & Schutze, 1999; Sinclair, 2004; Stubbs, 1995). Yoon (2016) used frequency and *mutual information* to identify the changes in target items over the course of one semester. Mutual information measures the strength of association between two or more words in a combination, showing how likely it is that these words will occur together. Researchers have proposed that mutual information be used to discover interesting collocations (Durrant & Schmitt, 2009; Siyanova-Chanturia, 2015; see the Method section for further information on mutual information). Yoon's (2016) longitudinal analysis showed no developmental changes in the essays, in either the narrative or argumentative genres, for the strength of association of verb-noun

combinations. Learner essays were further compared to L1 essays produced by 46 native speakers of English. The comparison of L1 and L2 essays suggested differences in L1 versus L2 combinations for association strength in argumentative essays (L2 combinations were not as strongly associated as L1 combinations), but not in narrative writing. Interestingly, Yoon found that all writers drew on a vast repertoire of high-frequency collocations, although L2 writers used three times fewer infrequent collocations ("sophisticated expressions," p. 54) compared to those used in L1 benchmark essays. The finding that the learners underused low-frequency collocations in writing was consistent with earlier research on L2 phrasal vocabulary (Durrant & Schmitt, 2009).

In another study employing a quantitative approach, Bestgen and Granger (2014) employed their novel CollGram technique to explore phraseological competence in the development of learner writing over the course of one semester by looking at L2 bigrams. The CollGram technique uses a large reference corpus (the Corpus of Contemporary American English in this case) to compute two association scores—mutual information and a *t* score—for each bigram found in a learner text. Similar to mutual information, the *t* score is a measure of association strength. Unlike mutual information, the *t* score correlates highly with corpus frequencies and, therefore, emphasizes high frequency, strongly associated collocations. Bestgen and Granger (2014) used as their learner corpus the Michigan State University Corpus of Second Language Writing that contains essays written by 57 ESL learners. The results of this longitudinal CollGram analysis of the learner corpus revealed a decrease over time in the use of two-word combinations made up of high-frequency words that are less typical of native speaker writing. The results of the pseudolongitudinal CollGram analysis revealed that the mean mutual information scores of the bigrams extracted from L2 texts were positively correlated with the perceived quality of L2 essays (that is, higher quality essays contained more strongly associated bigrams). In addition, Bestgen and Granger (2014) found a negative correlation between the quality of the essays and the bigrams that were absent in the reference corpus, meaning that poorer quality essays contained more non-attested bigrams.

Although English has been by far the most commonly investigated L2 in learner corpus research, scholars have also turned their attention to L2s other than English. Siyanova-Chanturia (2015) traced the use of Adj+N combinations in compositions written by 36 Chinese learners of Italian. Data collection took place over the course of 6 months, which was the duration of the language course the students attended. Siyanova-Chanturia (2015) looked both at the quality (how nativelike) and quantity (how many) of L2 combinations. Similar

to Yoon (2016) and Bestgen and Granger (2014), she adopted a frequency-based approach. Siyanova-Chanturia (2015) found that the number of Adj+N combinations was similar at the beginning versus the end of the intensive language course. However, she also reported that these learners were more likely to use higher frequency combinations in the essays written at the end of the 6-month period than in those written at the beginning. These high frequency combinations were also strongly associated collocations as suggested by mutual information scores (*città natale* "birth place," *aria fresca* "fresh air"). Thus, with time, Chinese learners of Italian were producing more targetlike (as suggested by corpora counts and association strength) collocations and fewer unidiomatic lower-frequency items. One downside of this (and other studies adopting a comparable quantitative approach) was that learner combinations were evaluated outside of the context in which they were originally used. If learners produce a frequent and/or strongly associated collocation, this does not necessarily imply that they know how to use it appropriately in context. Thus, ideally, quantitative analyses, such as those performed by Siyanova-Chanturia (2015) should be further supplemented by a qualitative exploration of the appropriateness of L2 combinations for that context.

Two further studies merit attention, not least because of their approach to data analysis. Garner and Crossley (2018) examined the development and use of bigrams and trigrams found in the spoken output of L2 English learners over a period of 4 months. Using latent curve modeling, which is a type of structural equation modeling, they found that the use of the target structures increased over the course of study. Critically, proficiency predicted the use of bigrams, but not trigrams, in that beginner L2 writers experienced greater growth, producing "more high frequency bigrams and more frequent bigrams" (p. 505) compared to advanced L2 users, who showed only marginal growth in their bigram use over the course of 4 months.

Finally, using a selection of essays from the Longitudinal Corpus of Chinese Learners of Italian (LOCCLI, Spina & Siyanova-Chanturia, 2018), Spina (2019) investigated the development of phraseological errors in beginner and intermediate learners of L2 Italian (L1 Chinese). In particular, she focused on two types of combinations, N+Adj and Adj+N,[2] produced over a period of 6 months. The LOCCLI has been annotated to allow users to identify errors, both grammatical (determiners, modifiers, agreement, number) and lexical (replacement of a lexical component, using a nonexistent combination, using an existing combination with a wrong meaning). Unlike the researchers who conducted the studies reviewed thus far, Spina (2019) used mixed-effects modeling, which allowed for a variety of predictors to be analyzed as fixed effects with

time being a key predictor. Spina found that L2 learners' use of word combinations was significantly affected by time, which interacted with proficiency. Although intermediate learners produced more errors toward the end of the course, beginner learners produced fewer errors over time. Interestingly, time also interacted with the type of word combination. Although the researcher observed a longitudinal decrease in Adj+N errors, N+Adj errors increased toward the end of the course.

Mixed-Effects Models in the Analysis of Learner Corpus Data

The learner corpus studies described above varied in a number of ways, such as the number of participants and the duration of the study period, multi-word expressions analyzed and the target L2, type of discourse and register studied, and other variables. What these longitudinal explorations had in common, however, was their approach to data analysis. With the exception of Spina (2019) and Garner and Crossley (2018), these studies either used traditional inferential statistical tests, such as the chi-square test of independence, log likelihood estimates, ANOVAs, *t* tests, correlation analysis, or descriptive linguistic and type-token analyses. Although these ways of data analysis have long been the accepted standard in applied linguistics, newer, more powerful and elegant techniques have recently gained ground. One of them, in particular, has quickly become the go-to analysis in experimental and, to a lesser, extent corpus research—mixed-effects models. Although mixed-effects modeling has become the gold standard in some linguistic subdisciplines such as psycholinguistics—for example, see the 2008 special issue of the *Journal of Memory and Language, 59*(4), its application in applied linguistics is still emergent.

Mixed-effects models present various advantages compared to means-based parametric statistical techniques such as ANOVAs and *t* tests (Cunnings, 2012; Cunnings & Finlayson, 2015; Gries, 2015; Linck & Cunnings, 2015; Murakami, 2016). For example, as Cunnings (2012) noted, the fixed effects component in a model can comfortably accommodate multiple independent variables, including categorical variables (males vs. females, L1 vs. L2), continuous predictors (age), as well as a mixture of both. Similarly, dependent variables can be continuous (reading or reaction times) or categorical (true or false). Further, and importantly for longitudinal research, mixed-effects models can be used to model change over time, whether the change is linear or not (Cunnings, 2012; Cunnings & Finlayson, 2015). In fact, researchers have long advocated for more powerful statistical tools to be used in the analysis of longitudinal L2 data (Ortega & Iberri-Shea, 2005). Additionally, in experimental research,

mixed-effects models can offer a more elegant (and statistically correct) alternative to running separate analysis by participants, where data are averaged across participants, and by items, where data are averaged across items. In mixed-effects models, participants and items are treated as crossed random effects and are included in a single analysis (Baayen, Davidson, & Bates, 2008; Cunnings, 2012). Mixed-effects models can also more easily handle missing data, which is an issue in longitudinal research, where attrition rates may pose a serious problem to the robustness of the analysis and results (Cunnings, 2012; Cunnings & Finlayson, 2015; Linck & Cunnings, 2015). All in all, mixed-effects modeling provides "a flexible and powerful tool for the analysis of a variety of data types" (Cunnings, 2012, p. 380).

One such data type is naturally corpus data. Ironically, Gries (2015) dubbed mixed-effects models "the most under-used statistical method in corpus linguistics" (p. 95), and, by extension, in learner corpus research. Only a handful of published studies have employed mixed-effects modeling for (various) longitudinal learner corpus data (Crossley, Skalicky, Kyle, & Monteiro, 2019; Crosthwaite & Jiang, 2017; Meunier & Littré, 2013; Murakami, 2016) and only one study to date has used mixed-effects models to explore L2 phraseological development over a period of time (Spina, 2019, see above). Echoing Cunnings (2012), Gries (2015) noted that corpus linguistics can benefit from mixed-effects models for the reasons similar to those for which psycholinguists have been using them for over a decade. But Gries (2015) went further, pointing out that corpus linguists may profit from mixed-effects models even more given the nature of corpus data—observational, unbalanced, and hence messy (much unlike carefully selected and controlled experimental stimuli used in psycholinguistic research).

The Present Study

The longitudinal explorations reviewed above varied in their duration from 4 months to three and a half years. Although, more prolonged periods may result in more and richer data, shorter periods can be highly informative too. For example, Crosthwaite and Jiang (2017) demonstrated the usefulness of short-term longitudinal corpus studies for the analysis of the development of stance expressions in L2 academic writing. Similarly, Siyanova-Chanturia (2015) showed that after only 6 months of an intensive course in the L2 country, the learners' use of N+Adj combinations in L2 Italian was more likely to be more native-like and idiomatic when compared to the output of the same learners at the beginning of the study period. In the present investigation, we sought to further demonstrate that key insights can be gained into the

complexities associated with the development of a L2 even if the period of study is relatively short.

The main aim of the present investigation was to track the development of phrasal vocabulary, that is of N+Adj word combinations, in L2 writing produced at two different points in time over the course of 6 months, relying on frequency of occurrence in the reference corpus and a selection of commonly used association measures. Thus, in our longitudinal learner corpus enquiry, we adopted a frequency-based approach (Hoey, 1991; Jones & Sinclair, 1974; Manning & Schutze, 1999), following the example of Durrant and Schmitt (2009), Lorenz (1999), Siyanova-Chanturia and Schmitt (2008), and others.

The gaps in the current learner corpus literature relevant to phrasal vocabulary motivated our study. First, although recent years have seen an increase in longitudinal studies looking at multi-word expression production in a L2, they are still relatively scarce. Indeed, Paquot and Granger (2012) and Laufer and Waldman (2011) called for a greater emphasis on research looking at the patterns of phraseological development over a period of time. And, as Bestgen and Granger (2014) further noted, although both longitudinal and cross-sectional studies can shed light on the development of the L2 phrasicon, only longitudinal explorations are able to track the development of the same individual learner over a period of time. It is, therefore, "essential to apply phraseological indices to truly longitudinal data" (Bestgen & Granger, 2014, p. 30).

Second, many of the currently available longitudinal studies are case studies employing a handful of participants (Crossley & Salsbury, 2011; Li & Schmitt, 2009, 2010; Li et al., 2014; Yuldashev et al., 2013). The larger-scale investigations have produced more conclusive and generalizable findings. Yet, even these studies have employed a limited number of L2 learners, ranging between 36 and 57 (Bestgen & Granger, 2014; Qi & Ding, 2011; Siyanova-Chanturia, 2015; Yoon, 2016).

Third, with some notable exceptions (Li et al., 2014; Myles et al., 1998; Siyanova-Chanturia, 2015), researchers have for the most part focused on upper intermediate or advanced learners. How the many and varied aspects of phrasal vocabulary use develop in less proficient learners has remained poorly understood (Granger & Bestgen, 2014).

Fourth, the majority of longitudinal learner corpus studies have looked at multi-word expressions in L2 English. Other L2s have so far been largely disregarded (but see Myles et al., 1998; Siyanova-Chanturia, 2015; Yuldashev et al., 2013). We believe it is important for a wider spectrum of L2s to be represented in learner corpus research and in the field of L2 acquisition more generally.

Finally, a number of researchers have recently voiced their concern about the statistical methods used to analyze learner corpus data (Gries, 2015). A strong argument has been advanced for the inadequacy of the currently used means-based statistical techniques such as ANOVAs and *t* tests. On the contrary, mixed-effects models have been shown to be much better suited for the analysis of a variety of data types including corpus data (Gries, 2015; Murakami, 2016).

In sum, to analyze N+Adj learner combinations based on different measures, the present corpus-based longitudinal exploration sought to advance the field of learner corpus research by: (a) employing a large pool of L2 learners ($n = 175$), (b) focusing on an underrepresented L2 (Italian), (c) looking at three different L2 proficiency levels (beginner, elementary, intermediate), and (d) using mixed-effects models, a flexible and powerful tool for corpus data analysis.

Method

Learner Corpus

To address the above gaps in learner corpus research, we employed the LOCCLI corpus (Spina & Siyanova-Chanturia, 2018; https://www.unistrapg.it/cqpwebnew/), a large-scale,[3] longitudinal, part-of-speech tagged corpus of L2 Italian. (The search tools are freely available following a straightforward account creation and login process). The part-of-speech tagging was carried out using TreeTagger (Schmid, 1994), specifically trained for native Italian texts. Although an evaluation of its accuracy with learner texts has not yet been conducted, a semiautomatic post-tagging revision ensures the reliability of the tagging process.

In total, 175 learners contributed two essays. Learners wrote one essay at the beginning of a 6-month intensive course of Italian, and the other at the end of the course. We used only two time points due to the logistics involved in having a large group of learners write essays in computer laboratories (each laboratory could accommodate only 40 learners, meaning multiple writing sessions took place). The learners were enrolled in a full-time course of Italian as a L2 that took place at a university in Italy. Students at three proficiency levels based on the Common European Framework of Reference for Languages (CEFR) contributed to the corpus: A1, or beginner ($n = 39$), A2, or elementary ($n = 86$), and B1, or intermediate ($n = 50$). All students came from the People's Republic of China and were between 17 and 33 years of age ($M = 20.5$ years, $SD = 2.7$; 105 females). On average, the students had spent 1.7 months in Italy (range: 0.5–5.0 months, $SD = 0.69$) prior to writing the first essay. The exact

Table 1 The size of the Longitudinal Corpus of Chinese Learners of Italian (LOCCLI) by data collection points in number of tokens for students at the three levels of the Common European Framework of Reference for Languages

	LOCCLI		
Data collection	A1[a]	A2[b]	B1[c]
Time 1	7,126	22,851	15,903
Time 2	9,487	24,117	17,386

[a]Beginner level.
[b]Elementary level.
[c]Intermediate level.

same 175 students who had written the first essay also wrote the second essay. We did not include in the corpus students who wrote only one of the two essays.

We offered participants three comparable essay topics: (a) "My first impression of Italy and Italians," (b) "My hobbies: What do I usually do in my free time?," and (c) "My last holidays." The rubric was given in the learners' L2 (Italian). We instructed the students not to write on the same topic more than once. Hence, all students chose two of the three topics.[4] The same group of teachers at the same university taught the participating students. These procedures helped to address factors such as topic, teaching style, and learning environment, which could potentially influence the nature of the corpus. The total size of the corpus was around 97,000 words. Table 1 shows the corpus size by proficiency group and by data collection point.

N+Adj Combinations

In this study, we focused on N+Adj combinations. Our decision to analyze N+Adj combinations was motivated by the observation that this Italian construction is extremely challenging for L2 learners (Spina, 2019) because nouns in Italian can be either preceded or followed by one or more adjectives.

Adjectival position is determined by syntactic and semantic constraints (Nespor, 1988). For example, an adjective follows the noun if it is modified by an adverb (*un film molto interessante* "a very interesting film") or by a complement (*un libro utile per gli studenti* "a useful book for the students"). Some adjectives precede the noun, as in *bel tempo* "nice weather," a sequence also allowed in Chinese. It is more common, however, for an adjective to follow the noun, as in *scuola elementare* "primary school."

We extracted the N+Adj combinations from the LOCCLI corpus through a simple search "noun followed by adjective," without pre-determining a span or specific syntactic relation between the two words. This resulted in 1,550 learner N+Adj combinations. Previous research has shown that the use of word combinations provided in a writing prompt can influence results (Staples, Egbert, Biber, & McClair, 2013). Consequently, we removed the 49 occurrences of the combinations *tempo libero* "free time," which was included in the prompt of composition Topic B, "My hobbies: What do I usually do in my free time?" We used only the remaining 1401 observations in our analysis. The data extracted from the corpus are available at https://www.iris-database.org/iris/app/home/detail?id=york:937023.

Measures Employed

In order to trace the acquisition of N+Adj combinations and to analyse the 1,401 observations produced by 175 Chinese learners of L2 Italian, we used five measures. Despite the existence of dozens of measures (Evert, 2005; Pecina, 2010), only a limited number has been used in the research on language learning. This small set has traditionally been based on two main dimensions of collocability: absolute frequency and strength of association, obtained by estimates of mutual information (Gablasova, Brezina, & McEnery, 2017).

Collocation measures are different ways of comparing the observed and the expected frequency values of word combinations, "putting different weight on different aspects of the collocational relationship" (Brezina, McEnery, & Wattam, 2015, p. 145). We considered the following dimensions of the collocational relationship in this study (each one corresponding to a different measure): frequency, exclusivity, type-token distribution, and directionality (Brezina et al., 2015). The five corresponding measures that we briefly describe below are: (a) phrase frequency, (b) mutual information, (c) lexical gravity, (d) delta $P_{forward}$, and (e) delta $P_{backward}$.

The fact that each of these measures differs from the others[5] and therefore is able to capture a unique aspect of the longitudinal development of lexical combinations in a L2 motivated our choice of these measures. Each of these unique aspects can provide a specific insight and allow for an accurate and comprehensive picture to emerge (Daudaravičius & Marcinkevičienė, 2004). In addition, although phrase frequency and mutual information are widely employed in L2 acquisition studies, the other three measures have rarely been used for learner data. As we have detailed in the following sections, we calculated the values for the selected measures using well established and commonly used equations.

Phrase Frequency

Phrase frequency accounts for overall repetition of word combinations (Gablasova et al., 2017), and represents an important indicator of their typicality (Brezina et al., 2015). According to Ellis (2002, 2012) and Ellis and Wulff (2015), the more learners are exposed to a given linguistic item such as a word combination, the stronger it is entrenched in their memory, and the easier it is accessed, processed, and produced.

Mutual Information

Mutual information has traditionally been described as the measure of association strength between two (or more) constituent words (Church & Hanks, 1990; Durrant & Schmitt, 2009). The strength of association is the result of the exclusivity of collocates, that is, "the extent to which the two words appear solely or predominantly in each other's company" (Gablasova et al., 2017, p. 160). The mutual information measure negatively correlates with frequency because it favors low-frequency word combinations. Because it emphasizes the exclusivity and the strength of the collocation relationship, it tends to assign higher scores to unusual and infrequent combinations. Because mutual information does not depend on the size of the corpus, Hunston (2002) argued that it is suitable for larger as well as smaller corpora. A mutual information score of 3 or above suggests a significant collocation threshold (Hunston, 2002; Stubbs, 1995). We obtained the values of mutual information using Equation 1 from Church and Hanks (1990, p. 77).

$$\mathrm{MI}\,(\mathrm{xy}) = \log_2 \frac{f\,(x, y)}{f\,(x)\,f\,(y)} \tag{1}$$

Lexical Gravity

Lexical gravity is a measure of diversification, based on type-token distribution (Daudaravičius & Marcinkevičienė, 2004; Gries, 2010). Higher lexical gravity values are obtained if the words included in a combination have a high type frequency value (Spina & Tanganelli, 2012), that is, if they are more diversified and compete for the slot close to the node word with other collocate types (Brezina et al., 2015). For example, *lingua italiana* ("Italian language") has a high lexical gravity value (12.54) in a reference corpus of Italian because both words may be included in a multitude of other word combinations, but *torre pendente* ("leaning tower") has a low lexical gravity value (0.005) because the adjective *pendente* occurs in very few other combinations. We calculated the values of

lexical gravity using Equation 2 from Daudaravičius and Marcinkevičienė's (2004, p. 331):

$$g(x) = \frac{f(x)}{n(x)} \; g'(y) = \frac{f(y)}{n'(y)} \quad (2)$$

Delta $P_{forward}$

Delta P scores account for the different ways in which one word attracts another word within a combination, and vice versa (Gries, 2013). The strength of the attraction between two words is not always symmetrical; traditional association measures often return high bidirectional associations for the two words, regardless of whether $Word_1$ selects $Word_2$ or $Word_2$ selects $Word_1$ (Gries, 2013, p. 149). Unlike mutual information and lexical gravity, delta P makes use of the full observed and expected contingency tables (Evert, Uhrig, Bartsch, & Proisl, 2017) and takes directionality into account, producing two different scores of collocational strength for each word combination. Delta $P_{forward}$ measures the extent to which "$word_1$ is much more predictive of $word_2$ than vice versa" (Gries, 2013, p. 148). For example, in *essere umano* ("human being"), the noun *essere* attracts the adjective *umano* more strongly than vice versa (delta $P_{forward}$: 0.45; delta $P_{backward}$: 0.12). We obtained the values of delta $P_{forward}$ using Equation 3 from Gries (2013, p. 144).

$$\Delta P_{forward} = f(word_2/word_1 = present) - f(word_2/word_1 = absent) \quad (3)$$

Delta $P_{backward}$

Delta $P_{backward}$ (Gries, 2013) measures the opposite, that is, the extent to which $Word_2$ more strongly predicts $Word_1$. For example, in *giorno feriale* ("working day"), the adjective *feriale* attracts the noun *giorno* more strongly than vice versa (delta $P_{forward}$: 0.002; delta $P_{backward}$: 0.63) because the noun is more frequent than the adjective and can be found in hundreds of other combinations, but the adjective *feriale* occurs almost exclusively with the noun *giorno*. We obtained the values of delta $P_{backward}$ using Equation 4 from Gries (2013, p. 144).

$$\Delta P_{backward} = f(word_1/word_2 = present) - f(word_1/word_2 = absent) \quad (4)$$

Reference Corpus

We based the values for the selected measures on text-external measures (Bestgen & Granger, 2014; Durrant & Schmitt, 2009), and, thus, we calculated them on the very large L1 Italian corpus *Paisà* (Lyding et al., 2014). The

Paisà corpus includes around 250 million tokens used in 380,000 written Italian texts extracted from the Web. The texts were selected among those licensed under Creative Commons (Attribution-ShareAlike and Attribution-Noncommercial-ShareAlike) and come mainly from Wikimedia Foundation sites and blog posts. The sources of the corpus, listed by Lyding et al. (2014), suggest that the vast majority of the texts included in the corpus are produced by native speakers of Italian. The *Paisà* corpus can, therefore, be considered to represent a variety of native Italian written language.

Although the use of L1 baseline (the native Italian corpus) to analyze L2 productions is not without criticism (e.g., see, Monteiro, Crossley, & Kyle, 2018; Ortega, 2016), the large size of the corpus, the variety of texts that it includes, and the range of language backgrounds among the writers of the texts make it a reliable benchmark for L2 written productions.

Analysis and Results

We used mixed-effects modeling for the reasons outlined above. We built five different models, in order to verify if and to what extent phrase frequency, association strength and exclusivity (mutual information), diversification according to type-token distribution (lexical gravity), and directionality (delta $P_{forward}$ and delta $P_{backward}$) varied as a function of time in combination with other variables.

A preliminary inspection of the data that the learners had produced showed that each of the five measures decreased at data collection Time 2, 6 months after writing the first essay at Time 1. Figure 1 shows plots for frequency, mutual information, and lexical gravity, the three measures where this decrease was strongest.

We built the five models using R (Version 3.5.1; R Core Team, 2018) and the R package lme4 (Version 1.1-18-1; Bates, Maechler, Bolker, & Walker, 2015). We used the packages sjPlot (Version 2.5.0; Lüdecke, 2018), ggplot2 (Version 3.0.0; Wickham, 2016), and yarrr (Version 0.1.5; Phillips, 2018) to build model plots and the MuMIn package (Version 1.42.1; Burnham & Anderson, 2002) for calculating R^2.

Each model had a different dependent variable, centered around the mean (Cunnings & Finlayson, 2015), corresponding to one of the five measures. Table 2 provides a summary of the dependent variables.

We selected the following predictors as fixed effects: time (data collection Time 1 and Time 2), proficiency (CEFR levels: A1, A2, B1), topic (the composition topics selected by the students among three different choices), the frequency and length of the first word (noun), the frequency and length of the second word (adjective), and a measure of lexical diversity. We calculated

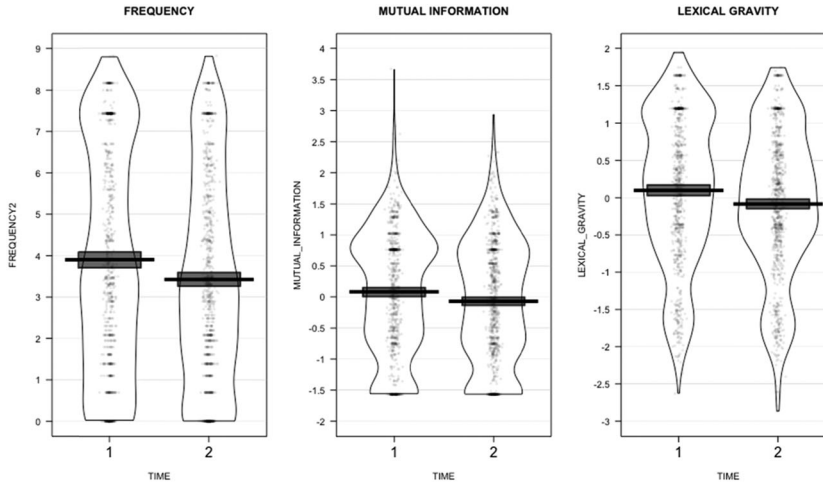

Figure 1 The longitudinal change in phrase frequency, mutual information, and lexical gravity scores.

Table 2 Summary of dependent variables used in the mixed-effects models (centered values shown in parentheses)

Variable	Range (centered)	*SD*	Median (centered)
Phrase frequency	0.00–6,788.00 (−0.38–7.43)	917.83	20.00 (−0.35)
Mutual information	−3.32–16.95 (−1.46–3.68)	3.86	2.87 (0.04)
Lexical gravity	−27.98–15.34 (−2.83–2.15)	8.99	−1.36 (0.05)
Delta $P_{forward}$	−0.001–0.49 (−0.34–18.68)	0.02	0.0006 (−0.28)
Delta $P_{backward}$	−0.001–0.64 (−0.34–17)	0.03	0.0007(−0.35)

lexical diversity using the Guiraud index (Guiraud, 1954, p. 53). The need to mitigate the effect of the nonequivalent lengths of the texts produced by L2 learners motivated this choice. According to Vermeer (2000), who compared different lexical diversity measures, the Guiraud index is appropriate, in particular, at early stages of vocabulary acquisition. We used Equation 5 to obtain the Guiraud index (R).

$$R = \frac{types}{\sqrt{tokens}} \quad (5)$$

We assigned a proficiency level individually to each participant through a placement test based on the CEFR scales. The test, developed at the University where the learners in question were enrolled, is routinely used with all learners

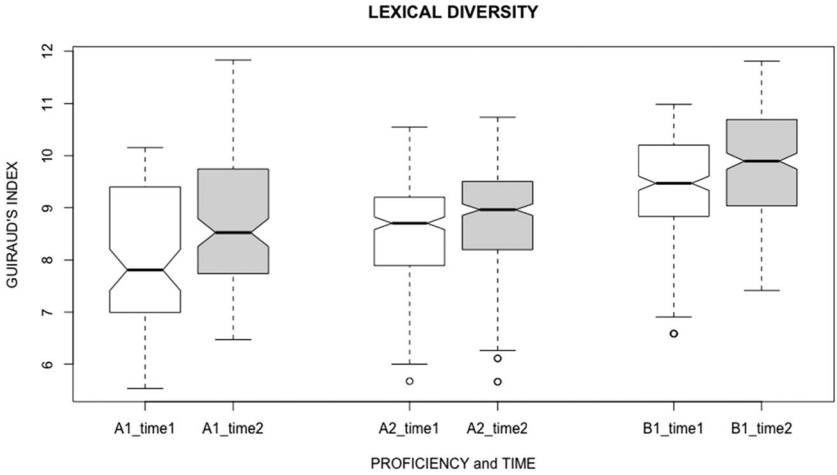

Figure 2 The development of lexical diversity across Time 1 (Data Collection 1) and Time 2 (Data Collection 2) for A1 (beginner), A2 (elementary), and B1 (intermediate) proficiency levels of the Common European Framework of Reference for Languages.

taking L2 Italian courses. The test scores are converted to a CEFR level to which students are assigned. The actual test scores—and thus reliability statistics—are not available. We, therefore, considered proficiency level to be an independent individual measure of proficiency and included it as a categorical variable (as the continuous data were not available) in the model as a fixed effect.

We included word length because earlier studies had shown that it could affect the acquisition process (Peters, 2016) as well as the online processing of multi-word expressions (Ellis, Simpson-Vlach, & Maynard, 2008). The length of constituent words has also been found to impact L1 and L2 speakers' judgements of collocation frequency (Siyanova-Chanturia & Spina, 2015). We could, thus, hypothesize that word length might affect the way in which the use of word combinations evolves over time.

We included lexical diversity as an indicator of lexical complexity (Paquot, 2018), capable of predicting lexical proficiency in language learners (Crossley, Salsbury, & McNamara, 2012, 2015). Visual inspection of the data extracted from the LOCCLI corpus showed that the essays at Time 2 had consistently higher lexical diversity indices compared to the essays at Time 1. The differences between Times 1 and 2 were significant across the three proficiency levels but were particularly prominent for A1 and B1 writers, as Figure 2 shows.

Table 3 Summary of the numeric independent variables used in the mixed-effects models (centered values shown in parentheses)

Variable	Range (centered)	*SD*	Median (centered)
Noun length	2.00–16.00 (−2.15–5.29)	1.81	5.00 (−0.02)
Adjective length	3.00–15.00 (−2.13–4.08)	1.86	6.00 (−0.06)
Noun frequency	22.00–579,082.00 (−0.79–7.11)	82,384.24	41,101.00 (−0.32)
Adjective frequency	0.00–306,784.00 (−0.90–3.61)	64,954.76	40,558.00 (−0.43)
Lexical diversity	5.53–11.83 (−2.92–2.31)	1.22	9.03 (0.01)

We centered the five numeric variables (lexical diversity and word—noun and adjective—length and frequency) around the mean. Table 3 provides a summary of the numeric independent variables and their centered values.

In order to model variation due to unsystematic individual differences in the use of N+Adj combinations, we used learners as a random effect, assuming different random intercepts for each of them. Although the 175 learners contributing to the corpus were taught by a group of teachers who had been part of the same team for a number of years, the effect of teaching style could not be excluded within the 17 different classes that the learners attended. Thus, the random effect of learners was nested within class. In addition, we used a by-subject random slope to model the repeated measures effect of time (Cunnings & Finlayson, 2015) because the learners' behavior might differ in their individual use of N+Adj combinations over time.

We adopted a stepwise approach for the model selection procedure following Gries (2015). We started with a model that contained the most comprehensive fixed-effect structure and first explored random effects to find the optimal random-effect structure, varying intercepts, and slopes for each of the three predictors. Once we had found the optimal random structure, we explored fixed effects to create the optimal fixed-effect structure. In this model selection process, we used a likelihood ratio test to compare pairs of models and to find the best fit (Baayen et al., 2008).

Fixed Effects and Interactions

Tables 4 to 8 summarize the five final models with phrase frequency, mutual information, lexical gravity, delta $P_{forward}$, and delta $P_{backward}$ as dependent variables.

The five models showed that time was a significant predictor of phrase frequency, lexical gravity, delta $P_{forward}$, and delta $P_{backward}$, affecting negatively each of these dependent variables. Time also had a negative estimate in terms of

Table 4 Fixed effects and interactions of Model 1 with phrase frequency as a dependent variable

Variable	Estimate	*SE*	*df*	*t*	*p* (>|*t*|)
(Intercept)	−0.49	0.15	1,387	−3.35	<.001
Noun length	0.12	0.03	1,387	3.52	<.001
Noun frequency	0.30	0.04	1,387	7.23	<.001
Adjective length	0.12	0.01	1,387	9.03	<.001
Adjective frequency	−0.97	0.10	1,387	−9.08	<.001
CEFR A2	−0.53	0.10	1,387	−4.93	<.001
CEFR B1	−0.51	0.11	1,387	−4.66	<.001
Composition Topic B[a]	0.23	0.06	1,387	3.53	<.001
Composition Topic C[b]	0.20	0.06	1,387	3.11	.002
Time 2	−0.46	0.12	1,387	−3.76	<.001
Noun length × Noun frequency	0.10	0.04	1,387	2.30	.020
Adjective length × Adjective frequency	0.17	0.01	1,387	11.38	<.001
CEFR A2 × Time 2	0.46	0.13	1,387	3.29	.001
CEFR B1 × Time 2	0.34	0.14	1,387	2.35	.018

Note. CEFR = Common European Framework of Reference for Languages; A2 = elementary level; B1 = intermediate level.
[a]"My hobbies: What do I usually do in my free time?"
[b]"My last holidays."

Table 5 Fixed effects and interactions of Model 2 with mutual information as a dependent variable

Variable	Estimate	*SE*	*df*	*t*	*p* (>|*t*|)
(Intercept)	−0.44	0.14	693.14	−3.25	.001
Noun length	0.06	0.02	1,389.83	2.40	.020
Adjective length	0.08	0.01	1,364.3	5.50	<.001
Adjective frequency	−0.81	0.11	1,362.76	−7.13	<.001
CEFR A2	−0.21	0.07	128.54	−2.89	.004
CEFR B1	−0.28	0.08	116.20	−3.67	<.001
Time 2	−0.08	0.05	1,249.28	−1.48	.130
Composition Topic B[a]	0.20	0.07	360.52	2.95	.003
Composition Topic C[b]	0.12	0.07	358.12	1.75	.080
Adjective length × Adjective frequency	0.09	0.02	1,356.26	5.57	<.001

Note. CEFR = Common European Framework of Reference for Languages; A2 = elementary level; B1 = intermediate level.
[a]"My hobbies: What do I usually do in my free time?"
[b]"My last holidays."

Table 6 Fixed effects and interactions of Model 3 with lexical gravity as a dependent variable

Variable	Estimate	*SE*	*df*	*t*	*p* (>|*t*|)
(Intercept)	0.30	0.07	200.14	4.07	<.001
Adjective length	0.11	0.02	1,381.19	4.25	<.001
Adjective frequency	0.08	0.02	1,392.95	3.26	.001
Noun frequency	0.21	0.02	1,392.83	8.08	<.001
CEFR A2	−0.22	0.07	144.37	−2.81	.005
CEFR B1	−0.35	0.08	125.83	−4.30	<.001
Time 2	−0.13	0.05	1,346.82	−2.51	.011
Adjective frequency × Noun frequency	0.05	0.02	1,383.15	2.53	.011

Note. CEFR = Common European Framework of Reference for Languages; A2 = elementary level; B1 = intermediate level.

Table 7 Fixed effects and interactions of Model 4 with delta $P_{forward}$ as a dependent variable

Variable	Estimate	*SE*	*df*	*t*	*p* (>|*t*|)
(Intercept)	0.60	0.10	110.77	5.85	<.001
Adjective frequency	0.06	0.02	1,365.17	2.41	.003
Noun frequency	−0.07	0.02	1,379.40	−2.90	.003
CEFR A2	−0.60	0.12	63.83	−4.96	<.001
CEFR B1	−0.69	0.12	27.19	−5.67	<.001
Time 2	−0.46	0.13	1,367.05	−3.47	<.001
CEFR A2 × Time 2	0.39	0.15	1,390.14	2.58	.009
CEFR B1 × Time 2	0.39	0.16	1,377.92	2.49	.012

Note. CEFR = Common European Framework of Reference for Languages; A2 = elementary level; B1 = intermediate level.

its effect on mutual information, but this estimate was not significant (Table 5). Additionally, for three dependent variables (frequency, delta $P_{forward}$, and delta $P_{backward}$) the effect of time interacted significantly with proficiency. These interactions showed that the negative effect of time on the three variables was not symmetric across the three levels of proficiency. Figures 3 to 5 show that the negative effect of time was stronger for Level A1 (beginner) learners. After 6 months of study, Chinese beginner learners of L2 Italian produced less frequent and less reciprocally attracting N+Adj combinations. The negative effect of time was found to be weaker for Level A2 (elementary) learners than the other two groups of learners. Interestingly, this negative effect turned positive for

Table 8 Fixed effects and interactions of Model 5 with delta $P_{backward}$ as a dependent variable

Variable	Estimate	*SE*	*df*	*t*	*p* (>|*t*|)
(Intercept)	0.26	0.11	1,391	2.32	.020
Adjective frequency	−0.15	0.02	1,391	−5.98	<.001
Noun frequency	0.20	0.02	1,391	7.88	<.001
CEFR A2	−0.38	0.11	1,391	−3.30	.001
CEFR B1	−0.43	0.12	1,391	−3.70	<.001
Time 2	−0.26	0.13	1,391	−1.94	.050
Composition Topic B[a]	−0.16	0.06	1,391	−2.33	.010
Composition Topic C[b]	0.04	0.07	1,391	0.63	.520
CEFR A2 × Time 2	0.27	0.15	1,391	1.85	.060
CEFR B1 × Time 2	0.32	0.16	1,391	2.04	.041

Note. CEFR = Common European Framework of Reference for Languages; A2 = elementary level; B1 = intermediate level.
[a]"My hobbies: What do I usually do in my free time?"
[b]"My last holidays."

Level B1 (intermediate) learners in the case of delta $P_{backward}$ (Figure 5). That is, after 6 months of study, Level B1 learners produced N+Adj combinations in which the adjective attracted more strongly the preceding noun than vice versa. This was the only positive effect of time among all the effects that we considered.

The topic of the compositions had a significant effect on frequency, mutual information, and delta $P_{backward}$. Other predictors that we found to be highly significant were noun and adjective frequencies. Both of these predictors were significant across the five models. Noun frequency had a positive effect on all of the measures except delta $P_{forward}$. Adjective frequency, on the other hand, had a positive effect on lexical gravity and delta $P_{forward}$, and a negative effect on the three remaining measures. As we expected, delta $P_{forward}$ values—where the noun more strongly predicts the adjective than vice versa—were higher for combinations with less frequent nouns (as in *essere umano* "human being"). Correspondingly, delta $P_{backward}$ values were higher for combinations with less frequent adjectives for the opposite reason. Infrequent adjectives are rarely found in a range of word combination and generally attract nouns more strongly than more frequent adjectives (as in *giorno feriale* "working day").

Noun and adjective lengths were also significant predictors, exhibiting highly symmetrical behavior. They had a positive effect on the same three

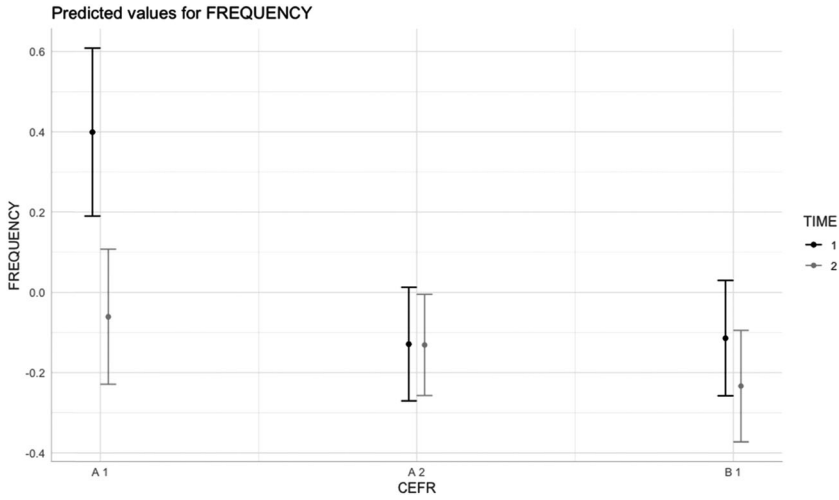

Figure 3 The interaction of time and proficiency effects on phrase frequency. CEFR = Common European Framework of Reference for Languages.

measures: phrase frequency, mutual information, and lexical gravity. The length of the two words, therefore, showed no significant effect on the directionality of the collocational relationship.

In addition, the models showed significant interactions between noun and adjective frequency and length in their effect on various dependent variables. Specifically, the two predictors of noun frequency and noun length interacted in their effect on phrase frequency. The length of the noun had less impact on phrase frequency when the noun was frequent. A similar interaction was found between adjective frequency and adjective length, where the adjective length had a weaker effect on phrase frequency and on mutual information when it was frequent. However, neither word length nor word frequency interacted significantly with time in its effect on the dependent variables. That is, the effects of word length and word frequency did not change over time.

Finally, we did not find the measure of lexical diversity significant in any of the five models, either as a fixed effect or in interaction with other variables.

Random Effects

One of the advantages of mixed-effects modeling compared to the more traditional statistical techniques such as ANOVA is the possibility of modeling variation due to unsystematic individual differences through random effects.

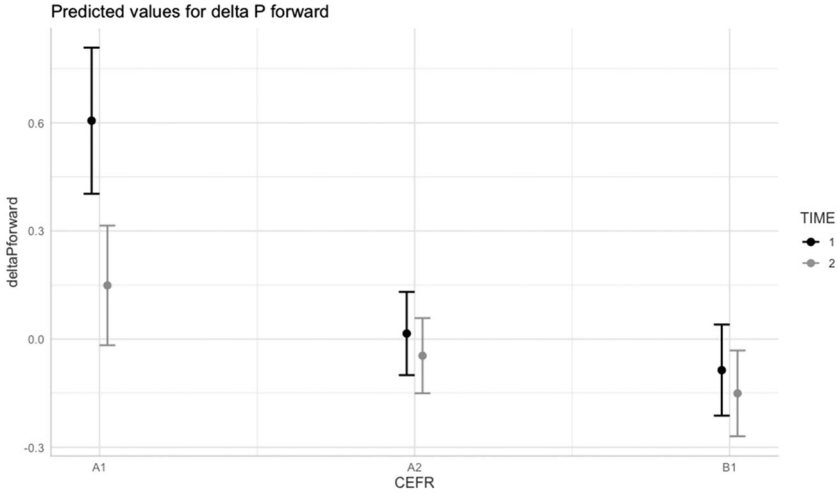

Figure 4 The interaction of time and proficiency effects on delta $P_{forward}$. CEFR = Common European Framework of Reference for Languages.

The final random effect structure for all five models, after the model selection procedure, included a single random effect where student was nested within class. The by-subject random slope of time was never significant. Table 9 shows the values related to the variance in the random effects for the final five models.

The analysis of the random effects of the five models revealed that there was little individual variance in the way the 175 learners produced N+Adj combinations within the 17 different classes. Across all the five dependent variables, learners' behavior in their production of N+Adj combinations did not exhibit significant variation and was fundamentally homogeneous. As an example, Figures 6 and 7 provide a visual representation of the random effects of learners nested within class on delta $P_{forward}$. Figure 6 shows the variation with respect to learners, and Figure 7 shows the variation with respect to class.

Table 10 provides the R^2 values for the five models (Hair, Black, Babin, & Anderson, 2013) that allowed us to determine the proportion of variance explained by each model. The values of marginal R^2 (R^2m), indicating the variance explained by the fixed effects alone, were very close to the conditional R^2 values (R^2c), indicating the variance explained by the whole model, including the random effects. In the case of phrase frequency and delta $P_{backward}$, the two values were the same. This observation confirmed that the learners' behavior

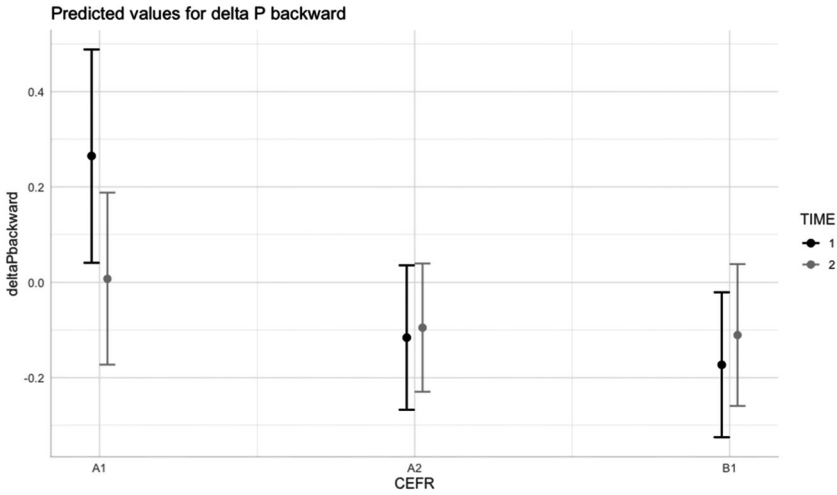

Figure 5 The interaction of time and proficiency effects on delta $P_{backward}$. CEFR = Common European Framework of Reference for Languages.

in the production of N+Adj combinations within the different classes was essentially homogeneous.

The fixed and random effects included in the models seemed to have much more predictive power in the case of phrase frequency, where they explained 21% of the variance in the dependent variable, compared to the other four models. This suggested that, at least for mutual information, lexical gravity, delta $P_{forward}$, and delta $P_{backward}$, the measures of collocability that we considered in this study had overall limited predictive power.

General Discussion

The main aim of the present large-scale longitudinal investigation was to examine the development of N+Adj word combinations in learner essays collected before and after a 6 month interval. In particular, we were interested in the change (if any) of learner configurations based on five measures: phrase frequency, mutual information, lexical gravity, delta $P_{forward}$, and delta $P_{backward}$. To this aim, we collected essays written by 175 learners of Italian (with L1 Chinese) of varying L2 proficiency at the beginning and at the end of an Italian as a L2 course. Given mounting calls for more complex, advanced, and rigorous ways of learner data analysis (Cunnings, 2012; Cunnings & Finlayson, 2015; Gries, 2015; Ortega & Iberri-Shea, 2005), we opted for mixed-effects modeling

Table 9 Random effect values for the dependent variables of the five models

Groups	Name	Variance	*SD*
Model 1: Phrase frequency			
Student × Class	(Intercept)	<0.0001	<0.0001
Class	(Intercept)	<0.0001	<0.0001
Residual		0.7913	0.8896
Model 2: Mutual information			
Student × Class	(Intercept)	0.001546	0.03932
Class	(Intercept)	0.000000	0.00000
Residual		0.912712	0.95536
Model 3: Lexical gravity			
Student × Class	(Intercept)	0.01228	0.1108
Class	(Intercept)	0.00000	0.0000
Residual		0.90850	0.9532
Model 4: Delta $P_{forward}$			
Student × Class	(Intercept)	0.0027387	0.05233
Class	(Intercept)	0.0005802	0.02409
Residual		0.9586447	0.97910
Model 5: Delta $P_{backward}$			
Student × Class	(Intercept)	<0.0001	<0.0001
Class	(Intercept)	<0.0001	<0.0001
Residual		0.9289	0.9638

where we fitted a separate model for each of the five measures considered. The following main findings emerged.

First, we found time (two data collection points) to be a negative predictor of phrase frequency, lexical gravity, delta $P_{forward}$, and delta $P_{backward}$. That is, we observed a decrease in the use of frequent combinations, combinations in which the first word strongly attracts the second word (delta $P_{forward}$), and combinations in which the second word strongly attracts the first word (delta $P_{backward}$) as a function of time. What this meant was that after 6 months, the students across the board (i.e., irrespective of the CEFR level) used a greater number of lower frequency phrasal configurations and combinations in which the two constituent words were not as strongly associated, mutually attracted,

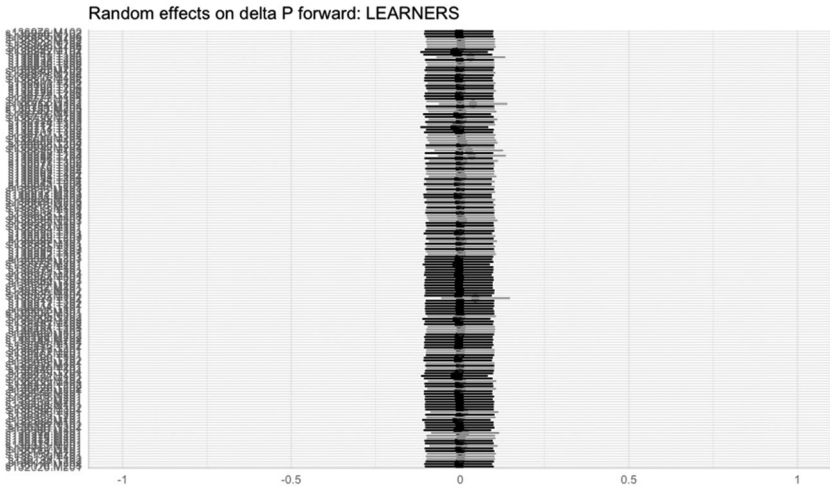

Figure 6 Plot of random effects on delta $P_{forward}$ with learners nested within class (Model 4).

Table 10 Marginal R^2 values (R^2m) and conditional R^2 values (R^2c) for the five models

Variable	R^2m	R^2c
Phrase frequency	.21445140	.21445140
Mutual information	.09112668	.09266328
Lexical gravity	.08377514	.09599840
Delta $P_{forward}$	.04330206	.04660277
Delta $P_{backward}$	.07657455	.07657455

or predictive of each other. It appears that as the learners' exposure to the target language and a variety of contexts increased, they started to use language more creatively and productively, experimenting with language. Examples of such unusual, creative (though grammatically correct) combinations that L1 speakers of Italian would not normally produce are *romanzo estero* "foreign novel," *verdura economica* "cheap/economic vegetables," and *nonno carino* "cute grandfather."

Our finding of time being a negative predictor of phrase frequency appears to contrast with the results reported in some of the earlier studies (Groom, 2009; Siyanova-Chanturia, 2015; also see Crossley & Salsbury, 2011). In a study methodologically comparable to the present investigation, Siyanova-Chanturia (2015) observed that 36 Chinese learners of L2 Italian used more, rather than

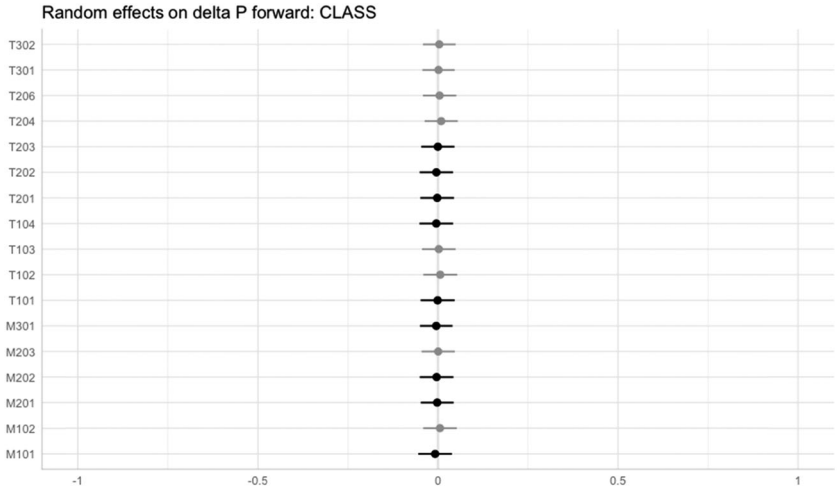

Figure 7 Plot of random effects on delta $P_{forward}$ with learners nested within class (Model 4).

fewer, frequent and strongly associated combinations at the end of an intensive language course and that the number of atypical word combinations (*studente calmo* "calm student," *spiaggia confortevole* "comfortable beach," *via affascinante* "fascinating street") decreased significantly. She concluded that a period of time as short as 5 months might be sufficient for L2 learners to begin to exhibit more targetlike, idiomatic phraseology. Similarly, using a cross-sectional paradigm, Groom (2009) found a positive correlation between more targetlike collocational usage and the time spent in a L2 country (1 month vs. 12 months).

Despite the differences that we observed in our investigation and in some of the earlier studies (Groom, 2009; Siyanova-Chanturia, 2015), our results support those of Bestgen and Granger (2014). In a study employing a longitudinal approach, these authors found a decrease in high-frequency collgrams but no differences in the number of low-frequency collgrams between the earlier and later written essays. Although we found a longitudinal decrease in higher frequency items, we observed no such differences for mutual information for the items extracted from the essays collected at the two points in time.

Our finding is further in line with Qi and Ding's (2011) longitudinal study. Although these researchers observed greater variation of L2 spoken phrases in Year 4 versus Year 1 production, many of these newly acquired combinations were used incorrectly. Qi and Ding (2011) concluded that as EFL learners become more capable language users and draw on a greater variety of L2 phrases,

they also err more frequently when using them. Although not directly comparable due to different methodological approaches adopted (quantitative in the present study vs. phraseological in Qi & Ding, 2011), the two studies have painted a complex picture of the development of L2 collocational knowledge. This is consistent with Laufer and Waldman's (2011) proposition that the development of collocation use is often slow and uneven. Based on the patterns of results that we observed, it is not unreasonable to expect L2 learners' phrasal production to become worse as a function of time before it slowly and gradually improves. Future large-scale studies with several data collection points conducted over an extended period of time (e.g., 2 to 3 years) and which also examine correctness and appropriacy are needed to be able to test this premise more directly.

Second, we found a decrease in the use of N+Adj combinations according to all measures that we considered as a function of proficiency. This suggests that more proficient learners (as per CEFR level) used fewer frequent, strongly associated, and diversified, as well as mutually attracted N+Adj configurations compared to those in the essays that less proficient writers produced. This was the case across the board, irrespective of the data collection period (beginning vs. end of the course). It appears that the more proficient Level B1 (intermediate) learners were more likely to experiment with a variety of nouns and adjectives to create less conventional (but grammatically correct) word combinations (*stile tranquillo* "tranquil style," *sera indimenticabile* "unforgettable evening," *sole biondo* "blond sun," *talento grande* "big talent") than their less proficient counterparts, the Level A1 (beginner) learners. This finding appears to align with some of the earlier learner corpus studies that investigated the role of proficiency in the acquisition and use of L2 word combinations. For example, Laufer and Waldman (2011) and Nesselhauf (2005) observed that more proficient EFL learners (or those with more years of English language learning) showed a proportion of incorrect collocations similar to that of less proficient learners (or those with fewer years of English language learning), or, indeed, produced more atypical, non-nativelike L2 word combinations than less proficient learners. Employing a qualitative, phraseological approach that drew on collocation dictionaries, L1 reference corpora, and L1 judgements, Nesselhauf (2005) found that up to one-third of L2 English V+N collocations (*make a decision*) that L1 German EFL writers produced deviated from L1 norms. Critically, the learners who had studied English between 10 and 17 years and those who had studied English between 5 and 10 years demonstrated a comparable proportion of collocations that deviated from L1 norms. This finding echoed Bahns and Eldaw's (1993) and other researchers' findings that

collocational competence does not always develop in parallel with general vocabulary knowledge. That is, although more proficient learners may possess larger vocabularies than their lower proficiency counterparts, their greater proficiency does not necessarily lead to higher accuracy or to idiomaticity, with L2 collocation.

Third, and, perhaps, more importantly, time (two data collection points) interacted with proficiency (CEFR level) in its effect on three of the five measures: frequency, delta $P_{forward}$, and delta $P_{backward}$. That is, time affected the three proficiency levels differently. Although, by and large, we found a longitudinal decrease in frequent and mutually attracted N+Adj combinations across all proficiency levels, the decrease was by far strongest for Level A1 (beginner) learners compared to the other groups. Our findings suggest that, although learners across all proficiency levels were more likely to start experimenting with unusual word combinations as their vocabularies grew over the course of 6 months, becoming less reliant on frequent N+Adj bigrams as their proficiency increased, the effect was most marked in these beginner (A1) learners.

This finding contrasts with the results reported in Garner and Crossley (2018). Looking at the development of bigrams and trigrams found in L2 spoken output, these authors found that the use of n-grams increased over the course of 4 months and that proficiency reliably predicted their use. Beginner learners demonstrated greater growth, producing a greater number of high frequency bigrams and, generally, using more frequent bigrams, compared to their advanced counterparts. The differences between the findings of the two studies could be due to the different modalities employed. In spoken production (e.g., Garner & Crossley, 2018), learners may become more reliant on phrase frequency, possibly due to the extra burden caused by time and communicative pressures, compared to written production (e.g., the present study).

The decrease in the learners' reliance on frequent and mutually attracted N+Adj combinations as a function of time, particularly evident in beginner writers, appears to also contradict the findings reported in another recent study by Spina (2019). Using a comparable albeit smaller data set, Spina (2019) found that beginner learners were the only group that produced fewer, rather than more, errors in Adj+N and N+Adj word combinations at the end of a 6-month period. However, Spina (2019) investigated a variety of both lexical and grammatical errors, which may not permit a useful comparison between the two studies. Overall, and independent of learner proficiency level, Spina (2019) observed a longitudinal decrease in Adj+N errors and a longitudinal increase in N+Adj errors, the latter agreeing with our observations.

Further, we also found constituent word length and word frequency to have a positive effect on the five measures. However, they never interacted with time, suggesting that no longitudinal changes in our measures as a function of word length and word frequency took place. Lexical diversity never affected the five measures, implying that vocabulary breadth did not appear to be linked to the development of N+Adj combinations over a period of time. Finally, we observed minimal individual variation among the L2 learners in their use of N+Adj combinations. But we did find that the use of frequent and strongly associated N+Adj combinations varied as a function of essay topic (L2 writers could choose from three topics: "My first impression of Italy and Italians," "My hobbies: What do I usually do in my free time?," and "My last holidays"). Because "topic can influence the strength of collocations as measured by the mutual information score" (Gablasova et al., 2017, p. 174), the inclusion of topic as a fixed effect allowed us to model and control this kind variation. The data showed, for example, that Topic B ("My hobbies: What do I usually do in my free time?") produced combinations with higher mutual information and phrase frequency values.

Overall, the present large-scale longitudinal exploration offers further evidence that multi-word expressions, in general, and N+Adj word combinations, in particular, pose difficulties for L2 learners. Importantly, our results point to the conclusion that the way in which L2 learners use N+Adj word combinations changes as a function of time, that is, as learners' proficiency increases. We found that more advanced learners produced more infrequent and weakly associated combinations (as suggested by the reference corpus) compared to beginner learners. It seems that, following extended exposure to the L2, as learners become more able language users and as they acquire more extensive (single word) vocabularies, they also tend to experiment more with word combinatorial mechanisms and, as a result, produce less idiomatic, less native-like word sequences. This finding may be attributed to the "inherent nature of collocations" (Laufer & Waldman, 2011, p. 665). As Laufer and Waldman (2011) noted, many collocations are semantically transparent (*do homework, make profit, strong wind, heavy rain*, *fast food, quick meal*) because their constituents are often high frequency words, likely to be known to L2 learners. This means that when learners encounter word combinations in their spoken or written input, they may deceptively perceive these word combinations to be easy or unworthy of attention. As a result, L2 learners simply fail to notice the many and varied phrasal configurations in their input and to attend to the fact that $Word_1$ appears with $Word_2$ and not with another, seemingly appropriate, synonymous $Word_3$. It may be that their experiences with the input are

insufficient to serve to preempt the *non-occurrence* of other combinations. Furthermore, producing appropriate L2 collocation has often been documented as challenging and difficult. It has been proposed that L2 learners fail to notice collocational relationships, paying attention to individual words and applying grammar rules to create novel utterances, rather than drawing on the highly formulaic, chunked, prefabricated nature of language (Sinclair, 1991; Wray, 2002).

Also clear from the results of the present study and from the literature reviewed is that L2 phraseological development is anything but straightforward to explain. Above, we have cited a wealth of studies and findings, some corroborative, others contradictory. Even where the nature of the learners' L1 and L2, the learning context, and the methodologies adopted were comparable, the results have proved to be different (e.g., the present study vs. Siyanova-Chanturia, 2015). What this complex evidence alludes to is that learning and using the myriads of phrasal configurations in a L2 is a mammoth task. It is a process that is fraught with difficulties and can be rather slow, a process in which more exposure and higher proficiency may not necessarily lead to a more idiomatic and targetlike output and may, indeed, result in lower levels of idiomaticity and greater reliance on lower frequency combinations whose constituent words are less likely to be associated and mutually attracted. The evidence that we have presented, considered in light of the results of the earlier studies, suggests that L2 learners' phrasal production may get worse as a function of time before it can slowly and gradually get better, further attesting to the highly complex and multifaceted process that is collocation learning.

Limitations and Future Research

Although the results of the present large-scale investigation have the potential to add to a better understanding of the development of phraseological competence over a period of time, there are still a number of unanswered questions that present fertile ground for future explorations. For example, the three longitudinal studies with the largest number of participants to date (the present study: 175 learners; Garner & Crossley, 2018: 57 learners; Siyanova-Chanturia, 2015: 36 learners) were conducted over a relatively short period of time (6, 4, and 6 months, respectively). Future large-scale explorations should be conducted over longer periods of time (e.g., 2 to 3 years) with a greater number of data collection points (e.g., two-three per year).

Arguably, one of the limitations of the present study is the relatively small corpus (around 97,000 words). Although this shortcoming has not jeopardized the analyses that we conducted or the findings that we have reported, future

studies should aim either to collect more essays (employ more participants) or require learners to produce longer essays or pieces of spoken discourse (proficiency permitting). When working with corpora, size does matter, particularly given the variability in L2 writing. Further, although logistically challenging, longitudinal analyses of spoken corpora with more than just a handful of participants (akin to Garner & Crossley, 2018) should take center stage in learner corpus research. Writing—as a task and activity—is off-line, in the sense that learners can take time to formulate their ideas and can delete, correct, or rewrite. Spoken discourse is online, in the sense that it often happens under time pressure with little or no time to prepare and limited opportunity to self-correct. The investigation and comparison of the two modalities (writing vs. speaking) is likely to reveal some interesting similarities as well as differences. Critically, future studies should adopt advanced and rigorous methods of data analysis that are able to explain the variance in L2 data (between learners) and to track collocation development over time (such as mixed effects models). In fact, it is possible that some of the dissimilarities in the developmental patterns that we observed and those in the earlier research can be partly explained by the different statistical approaches adopted. Finally, in line with Ortega (2009), we believe that a wider range of L2s represented in longitudinal studies—and learner corpus research, in general—is bound to enrich and further advance the field.

Conclusion

In conclusion, we have sought to provide in our learner corpus exploration a detailed account of L2 phraseological development—with a focus on N+Adj word combinations—over a 6-month period. With 175 learners contributing one essay each at the beginning and at the end of an Italian as a L2 course, it is the largest-scale longitudinal investigation available to date. Crucially, the use of five measures—phrase frequency, mutual information, lexical gravity, delta $P_{forward}$, and delta $P_{backward}$—allowed for a more holistic and complex picture to emerge. Finally, this study contributes to a growing trend in the field of applied linguistics, in general, and learner corpus research, in particular, toward more powerful methods of data analysis. Mixed effects models allow for a variety of categorical and numerical predictors, as well as their interactions, to be analyzed in a single model. Given the nature of corpus data—observational, unbalanced, and often messy (Gries, 2015)—mixed effects modeling has the potential to offer the most complex and complete account of L2 developmental patterns that the more traditional ways of data analysis may be unable to capture.

Final revised version accepted 22 August 2019

Notes

1 The main focus of the current investigation was on longitudinal learner corpus research. Thus, the literature reviewed pertains to studies conducted over a period of time. For an exhaustive overview of learner corpus research in the context of multi-word expressions, we direct the interested reader to Paquot and Granger (2012) and Granger (2019).

2 Most adjectives in Italian follow the noun that they modify (*piazza grande* "main square," *tempo libero* "free time," *libro interessante* "interesting book"; see Nespor, 1988). Some adjectives, however, appear before the noun that they modify (*bel tempo* "good weather," *buon amico* "good friend"). Depending on the meaning, some adjectives may appear before or after the noun (*vestito caro* "expensive dress," *cara sorella* "dear sister"). This variation in adjectival position is a common source of confusion and errors for learners of L2 Italian (Spina, 2019).

3 Large-scale in terms of the number of learners who participated in the study. Although the resulting corpus was relatively small, the sizable participant pool ($N = 175$) by far surpassed current longitudinal learner-corpus studies.

4 The three topics were distributed as follows in the 350 learner compositions: topic a = 57 (A1, Time 1 = 7; A1, Time 2 = 5; A2, Time 1 = 12; A2, Time 2 = 16; B1, Time 1 = 10; B1, Time 2 = 7); topic b = 160 (A1, Time 1 = 26; A1, Time 2 = 10; A2, Time 1 = 53; A2, Time 2 = 26; B1, Time 1 = 25; B1, Time 2 = 20); and topic c = 133 (A1, Time 1 = 6; A1, Time 2 = 24; A2, Time 1 = 21; A2, Time 2 = 44; B1, Time 1 = 15; B1, Time 2 = 23).

5 Of note is that we did not use *t* score, a measure of association strength traditionally used in L2 acquisition research, in this study. The *t* score tends to highlight frequent combinations (see Durrant & Schmitt, 2009), thus overlapping with phrase frequency in what it can reveal about a given word combination.

Open Research Badges

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://osf.io/tvyxz/wiki.

References

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

Arnon, I., & Christiansen, M. H. (2017). The role of multiword building blocks in explaining L1–L2 differences. *Topics in Cognitive Science*, *9*, 621–636. https://doi.org/10.1111/tops.12271

Arnon, I., McCauley, S. M., & Christiansen, M. H. (2017). Digging up the building blocks of language: Age-of-acquisition effects for multiword phrases. *Journal of Memory and Language*, *92*, 265–280. https://doi.org/10.1016/j.jml.2016.07.004

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*, 67–82. https://doi.org/10.1016/j.jml.2009.09.005

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System*, *21*, 101–114. https://doi.org/10.1016/0346-251X(93)90010-E

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. Retrieved from https://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

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

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer-Verlag.

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

Christiansen, M. H., & Chater, N. (1999). Connectionist natural language processing: The state of the art. *Cognitive Science*, *23*, 417–437. https://doi.org/10.1016/S0364-0213(99)00010-5

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Crossley, S., & Salsbury, T. L. (2011). The development of lexical bundle accuracy and production in English second language speakers. *IRAL-International Review of Applied Linguistics in Language Teaching*, *49*(1), 1–26. https://doi.org/10.1515/iral.2011.001

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*, 243–263. https://doi.org/10.1177/0265532211419331

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*, 570–590. https://doi.org/10.1093/applin/amt056

Crossley, S. A., Skalicky, S., Kyle, K., & Monteiro, K. (2019). Absolute frequency effects in second language lexical acquisition. *Studies in Second Language Acquisition*, *41*, 721–744.

Crosthwaite, P., & Jiang, K. (2017). Does EAP affect written L2 academic stance? A longitudinal learner corpus study. *System*, *69*, 92–107. https://doi.org/10.1016/j.system.2017.06.010

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*, 369–382. https://doi.org/10.1177/0267658312443651

Cunnings, I., & Finlayson, I. (2015). Mixed effects modeling and longitudinal data analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 159–181). New York, NY: Routledge.

Daudaravičius, V., & Marcinkevičienė, R. (2004). Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics*, *9*, 321–348. https://doi.org/10.1075/ijcl.9.2.08dau

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL), New Series*, *2*, 225–246.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL-International Review of Applied Linguistics in Language Teaching*, *47*(2), 157–177. 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. (2012). What can we count in language, and what counts in language acquisition, cognition, and use? In S. Gries & D. Divjak (Eds.), *Frequency effects in language learning and processing* (pp. 7–34). Berlin, Germany: Mouton de Gruyter.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, *42*, 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

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

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, *33*, 1–36. Retrieved from https://10.1111/j.1551-6709.2009.01023.x

Evert, S. (2005). *The statistics of word co-occurrences: Word pairs and collocations*. (Unpublished doctoral dissertation). Stuttgart, Germany: Institut fur maschinelle Sprachverarbeitung, University of Stuttgart.

Evert, S., Uhrig, P., Bartsch, S., & Proisl, T. (2017). E-VIEW-alation—A large-scale evaluation study of association measures for collocation identification. In K. Isok,

T. Carole, J. Miloš, K. Jelena, K. Simon, & B. Vít (Eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2017 conference* (pp. 531–549). Brno, Czech Republic: Lexical Computing. Retrieved from https://elex.link/elex2017/wp-content/uploads/2017/09/paper32.pdf

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching, and testing* (pp. 75–94). Harlow, UK: Longman.

Foster, P., Bolibaugh, C., & Kotula, A. (2014). Knowledge of nativelike selections in a L2. *Studies in Second Language Acquisition*, *36*, 101–132. https://doi.org/10.1017/S0272263113000624

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

Garner, J., & Crossley, S. A. (2018). A latent curve model approach to studying L2 n-gram development. *The Modern Language Journal*, *102*, 494–511.

Gilquin, G. (2007). To err is not all. What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift fur Anglistik und Amerikanistik*, *55*, 273–291.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 145–160). Oxford, UK: Oxford University Press.

Granger, S. (2019). Formulaic language in learner corpora. Collocations and lexical bundles. In A. Siyanova-Chanturia & A. Pellicer-Sanchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 228–247). London, UK: Routledge.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, *52*(3), 229–252. https://doi.org/10.1515/iral-2014-0011

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge handbook of learner corpus research*. Cambridge, UK: Cambridge University Press.

Granger, S., & Paquot, M. (2009). Lexical verbs in academic discourse: A corpus-driven study of learner use. In M. Charles, D. Pecorari, & S. Hunston (Eds.), *Academic writing: At the interface of corpus and discourse* (pp. 193–214). New York, NY: Continuum.

Gries, S. T. (2010). Useful statistics for corpus linguistics. In A. Sanchez & M. Almela (Eds.), *A mosaic of corpus linguistics: Selected approaches* (pp. 269–291). Frankfurt am Main, Germany: Peter Lang.

Gries, S. T. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics*, *18*(1), 137–166. https://doi.org/10.1075/ijcl.18.1.09gri

Gries, S. T. (2015). Statistics for learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 159–181). Cambridge, UK: Cambridge University Press.

Groom, N. (2009). Effects of second language immersion on second language collocational development. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language* (pp. 21–33). London, UK: Palgrave Macmillan.

Guiraud, P. (1954). *Les Charactères Statistiques du Vocabulaire. Essai de méthodologie*. Paris: Presses Universitaires de France.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2013). *Multivariate data analysis*. Harlow, UK: Pearson Education.

Henderson, A., & Barr, R. (2010). Comparing indicators of authorial stance in psychology students' writing and published research articles. *Journal of Writing Research*, *2*, 245–264.

Hoey, M. (1991). *Patterns of lexis in text*. Oxford, UK: Oxford University Press.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London, UK: Routledge.

Howarth, P. (1998). The phraseology of learners' academic writing. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and application* (pp. 161–186). Oxford, UK: Oxford University Press.

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

Jackendoff, R. (1995). The boundaries of the lexicon. In M. Everaert, E. van der Linded, A. Schenk, & R. Schreuder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 133–165). Hillside, NJ: Erlbaum.

Jones, S., & Sinclair, J. M. (1974). English lexical collocations: A study in computational linguistics. *Cahiers de lexicologie*, *24*, 15–61.

Langacker, R. (1987). *Foundations of cognitive grammar* (Vol. *1*). Stanford, CA: Stanford University Press.

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

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*, 387–417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing*, *18*(2), 85–102. https://doi.org/10.1016/j.jslw.2009.02.001

Li, J., & Schmitt, N. (2010). The development of collocation use in academic texts by advanced L2 learners: A multiple case study approach. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 22–46). New York, NY: Bloomsbury.

Li, P., Eskildsen, S. W., & Cadierno, T. (2014). Tracing an L2 learner's motion constructions over time: A usage-based classroom investigation. *The Modern Language Journal*, *98*, 612–628.

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

Lorenz, G. (1999). *Adjective intensification – Learners versus native speakers. A corpus study of argumentative writing*. Amsterdam, Netherlands: Rodopi.

Lüdecke, D. (2018). sjPlot: Data visualization for statistics in social science [Computer software]. Retrieved from https://zenodo.org/record/2400856

Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., … Pirrelli, V. (2014). The PAISÀ corpus of Italian Web texts. In F. Bildhauer & R. Schäfer (Eds.), *Proceedings of the 9th Web as corpus workshop (WaC-9)* (pp. 36–43). Gothenburg, Sweden: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W14-0406

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

McCauley, S. M., & Christiansen, M. H. (2017). Computational investigations of multiword chunks in language learning. *Topics in Cognitive Science*, *9*, 637–652. https://doi.org/10.1111/tops.12258

Meunier, F., & Littré, D. (2013). Tracking learners' progress: Adopting a dual "corpus cum experimental data" approach. *The Modern Language Journal*, *97*, 61–76. https://doi.org/10.1111/j.1540-4781.2012.01424.x

Monteiro, K. R., Crossley, S. A., & Kyle, K. (2018). In search of new benchmarks: Using L2 lexical frequency and contextual diversity indices to assess second language writing. *Applied Linguistics*, Advance online publication. https://doi.org/10.1093/applin/amy056

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

Myles, F., Hooper, J., & Mitchell, R. (1998). Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning*, *48*, 323–364.

Nespor, M. (1988). Il sintagma aggettivale. In L. Renzi, G. Salvi, & A. Cardinaletti (Eds.), *Grande grammatica italiana di consultazione* (pp. 439–455). Bologna, Italy: Il Mulino.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, *24*, 223–242. https://doi.org/10.1093/applin/24.2.223

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

Ortega, L. (2009). *Understanding second language acquisition*. London, UK: Hodder.

Ortega, L. (2016). Multi-competence in second language acquisition: Inroads into the mainstream? In V. Cook & L. Wei (Eds.), *The Cambridge handbook of linguistic multi-competence* (pp. 50–76). Cambridge, UK: Cambridge University Press.

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 useful toolbox to delimitate CEFR levels in higher education? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly*, *15*, 29–43.

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

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, *44*(1-2), 137–158.

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

Phillips, N. D. (2018). *YaRrr! The pirate's guide to R*. Retrieved from https://bookdown.org/ndphillips/YaRrr/

Qi, Y., & Ding, Y. (2011). Use of formulaic sequences in monologues of Chinese EFL learners. *System*, *39*(2), 164–174. https://doi.org/10.1016/j.system.2011.02.003

R Core Team (2018). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK. Retrieved from https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf

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

Sinclair, J. M. (2004). *Trust the text: Language, corpus and discourse*. London, UK: Routledge.

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., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review*, *64*, 429–458.

Siyanova-Chanturia, A., & Spina, S. (2015). Investigation of native speaker and second language learner intuition of collocation frequency. *Language Learning*, *65*, 533–562. https://doi.org/10.1111/lang.12125

Siyanova-Chanturia, A., & Van Lancker Sidtis, D. (2019). What on-line processing tells us about formulaic language. In A. Siyanova-Chanturia & A. Pellicer-Sanchez

(Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 38–61). London, New York: Routledge.

Spina, S. (2019). The development of phraseological errors in Chinese learner Italian: A longitudinal study. In A. Abel, A. Glaznieks, V. Lyding, & L. Nicolas (Eds.), *Widening the scope of learner corpus research. Selected papers from the fourth Learner Corpus Research Conference* (pp. 95–119). Louvain, Belgium: Presses Universitaires de Louvain.

Spina, S., & Siyanova-Chanturia, A. (2018) The Longitudinal Corpus of Chinese Learners of Italian (LOCCLI). Poster presented at the *13th Teaching and Language Corpora* conference, University of Cambridge, UK.

Spina, S., & Tanganelli, E. (2012). Collocations as an index for distinguishing text genres. *Corpus*, *11*, 73–89.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, *12*(3), 214–225. https://doi.org/10.1016/j.jeap.2013.05.002

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, *2*(1), 23–55. https://doi.org/10.1075/fol.2.1.03stu

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*, 65–83. https://doi.org/10.1177/026553220001700103

Waibel, B. (2008). *Phrasal verbs: German and Italian learners of English compared*. Saarbrucken, Germany: VDM.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag.

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

Wray, A. (2019). Concluding question. Why don't second language learners more proactively target formulaic sequences? In A. Siyanova-Chanturia & A. Pellicer-Sanchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 248–269). London, UK: Routledge.

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

Yuldashev, A., Fernandez, J., & Thorne, S. L. (2013). Second language learners' contiguous and discontiguous multi-word unit use over time. *The Modern Language Journal*, *97*(S1), 31–45.

Appendix: Accessible Summary (also publicly available at https://oasis-database.org)

Investigating Noun+Adjective Word Combinations in Second Language Writing

What This Research Was About and Why It Is Important

Language is known to comprise a wide range and very high number of word combinations: small sets of words that very often occur together. Despite the considerable interest that word combinations have received in the recent years from researchers, particularly in the context of learner spoken and written discourse, we still know very little about the development of such combinations over a period of time. The present longitudinal study sought to track the development of noun+adjective combinations in essays produced by second language learners of Italian with Chinese as their first language.

What the Researchers Did

- We collected learner essays produced at two different points in time, at the beginning and end of an intensive 6-month-long language course in Italy.
- We employed a large pool of second language learners ($n = 175$) of three proficiency levels: beginner, elementary, and intermediate (measured according to the Common European Framework of Reference).
- We analyzed learners' word combinations (nouns + adjectives) using five measures of how common, nativelike, and natural these combinations were (using data about word combinations from a corpus of Italian).

What the Researchers Found

- We found that after 6 months, students across all proficiency levels started to use language more creatively and productively, producing combinations that deviated more from native-speaker norms than at the start.
- We also found that time affected the three proficiency levels differently. Although we found a decrease over time in frequent and nativelike noun+adjective combinations across all proficiency levels, the decrease was strongest for beginner learners compared to the other groups.
- Our findings suggest that while learners across all proficiency levels were more likely to start experimenting with unusual word combinations as their proficiencies grew over the course of 6 months, the effect was most marked in the beginner learners.

Things to Consider

- Our findings suggest a complex picture, wherein higher language proficiency and greater exposure to the target language do not necessarily entail more native-speaker-like output, and may, in fact, result in greater reliance on less natural, lower frequency combinations.
- Future large-scale explorations should be conducted over longer periods (e.g., 2 to 3 years) with more data collection points (e.g., three per year), and examine appropriacy and accuracy of word combinations.
- Future studies should aim either to collect more essays (i.e., employ more participants) or require learners to produce longer essays or spoken discourse (proficiency permitting).
- Finally, while the present study has looked at written discourse, future longitudinal research should further focus on large-scale explorations of second language word combinations in spoken discourse.

Materials and Data: Data and search tools are publicly available at https://www.iris-database.org/iris/app/home/detail?id=york:937023 and https://www.unistrapg.it/cqpwebnew/

How to cite this summary: Siyanova-Chanturia, A., & Spina, S. (2019). Investigating noun+adjective word combinations in second language writing. *OASIS Summary* of Siyanova-Chanturia, A. & Spina, S. in *Language Learning*. https://oasis-database.org

This summary has a CC BY-NC-SA license.