Using Syntactic Co-occurrences to Trace Phraseological Complexity Development in Learner Writing: Verb + Object Structures in LONGDALE

Magali Paquot, Hubert Naets, and Stefan Th. Gries

1 Introduction

Many recent studies of word combinations in learner writing have relied on the use of statistical collocations to assess English as a Foreign Language (EFL) learners' phraseological competence.[1] Statistical collocations are word combinations such as *severe* + *weather*, *take* + *time*, and *ride* + *horse* that "co-occur more often than their respective frequencies and the length of text in which they appear would predict" (Jones & Sinclair 1974, 19). In learner corpus research (LCR), statistical collocations have typically been identified by means of association measures such as the pointwise mutual information (MI) score and the t-score (Gablasova et al. 2017).[2] The two statistical measures have often been used in tandem on the ground that MI will rank best word combinations made up of low-frequency words (e.g. *substantiating evidence*, *corroborative evidence*) while t-score will give prominence to word combinations composed of high-frequency words (e.g. *further evidence*, *empirical*

[1] This chapter is based on a paper first presented by the first two authors at the 3rd Learner Corpus Research Conference, the Netherlands, September 11–13, 2015.

[2] Note that there is also an extensive body of research that has approached word combinations in the form of constructions, most particularly verb-argument constructions, in learner language (e.g. Gries & Wulff 2005; Ellis et al. 2016; Kyle & Crossley 2017). In this particular strand of research, and as noted by one reviewer, other (directional) association measures such as Delta P have been used and promoted to investigate the degree of attraction of a lemma to a slot in one particular construction or the preference of a lemma for one particular construction over another. In this study, we did not use Delta P because we wanted to situate our work against previous research on collocations in learner writing, which is nearly exclusively based on t- and MI-scores.

evidence) (Durrant & Schmitt 2009).[3] The following procedure has often been adopted to assess collocations in a learner corpus. First, word pairs or co-occurrences are extracted from a large reference corpus such as the *British National Corpus* (BNC) and association measures are computed for each of them. Second, co-occurrences are extracted from learner corpora and the corresponding BNC-derived association measures are assigned to learners' production. Third, for each learner text, a mean MI score and/or t-score is computed (e.g. Bestgen & Granger 2014; Paquot 2019) or co-occurrences are categorized into collocational bands (see Granger & Bestgen 2014 for more details on this specific procedure).

This approach has generated a wealth of interesting results that stress the value of a frequency-based approach to phraseology (cf. Granger & Paquot 2008) to assess foreign language proficiency and trace foreign language development. Durrant and Schmitt (2009), for example, showed that, compared to native writers, L2 writers of English tend to overuse high-frequency noun/adjective + noun pairs identified by high t-scores (e.g. *good example*, *long way*, *hard work*) but underuse less common, strongly associated collocations as identified by high MI scores (e.g. *densely populated*, *bated breath*, *preconceived notions*). In a study that investigated the full range of contiguous word pairs instead of being restricted to modifier + noun sequences, Granger and Bestgen (2014) demonstrated that the same difference can be observed between intermediate and advanced EFL learners in a subset of learner texts from the *International Corpus of Learner English* (ICLE; Granger et al. 2009): advanced learners have a lower proportion of high-frequency collocations (t-score) and a higher proportion of lower-frequency collocations (MI score).

Bestgen and Granger also reported a significant decrease in the use of collocations made up of high-frequency words (average t-score values) from time 1 to time 3 (a six-month difference) in the *Michigan State University Corpus of Second Language Writing*, but no effect of time on average MI scores was found, which the authors argued can be explained by "the low frequency of the bigrams in the learners' input coupled with the short period in time covered" (2014, 37). However, mean MI scores of the bigrams used by L2 writers were shown to be positively correlated with the quality of the essays, while there was a negative correlation between the quality of the texts and the proportion of bigrams that were absent in the reference corpus,

[3] This "dichotomous description" of collocation use, however, has also recently been criticized as "too general to be useful in [Language Learning Research]" (Gablasova et al. 2017, 163).

most of which were shown to be erroneous. In a recently published follow-up study, Bestgen and Granger (2018) investigated the development of collocational strength in bigrams (types and tokens) in a set of essays produced by French-speaking learners of English and collected at a two-year interval within the framework of the *Longitudinal Database of Learner English* project (LONGDALE; see Section 2). Among the many results, the analyses based on tokens revealed a significant increase of bigrams with low t-score (Cohen's d = 0.36) and a significant decrease of bigrams with high t-score in the second set of essays (Cohen's d = 0.28). By contrast, only the category of high MI bigrams (types and tokens) showed a significant increase with time, with medium to large effect sizes (Cohen's d of 0.68 and 0.78, respectively).[4]

All the studies above rely on a positional or surface model of co-occurrence, where words are said to co-occur when they appear within a close distance from each other, measured in number of intervening words (cf. Evert 2005, 18–19; Evert 2008). More precisely, the above studies all set a minimal collocational span and investigated statistical co-occurrences as bigrams, i.e. contiguous sequences of two words. A significant advantage of positional co-occurrences and bigrams for L2 research (and more particularly for research with applied perspectives in language teaching and assessment, cf. Bestgen 2017) is that they can be extracted easily and quickly from corpora and their frequencies can be measured reliably with fully automatic techniques. The downside, however, is that positional co-occurrences are blind to syntactic relations such as subject + verb, verb + object, predicative adjectives, verb + particle or the pattern N *of* N in English. Some of these relations, however, have been shown to be particularly problematic for EFL learners, even at an intermediate to advanced level (e.g. Nesselhauf 2005, or Laufer & Waldman 2011, on verb + noun combinations).

As a consequence, in a study that aimed to put forward and operationalize the construct of phraseological complexity (see below for more information), Paquot (2019) adopted a relational or syntactic model of co-occurrences to investigate EFL learners' use of adjective modifiers, adverb modifiers, and verb + noun structures at the B2, C1, and C2 proficiency levels of the Common European Framework of References for Languages (CEFR; Council of Europe 2001) in the

[4] Unfortunately, as the method used in Bestgen and Granger (2018) is not the same as that used in Bestgen and Granger (2014), it is not straightforward to compare the six-month vs. two-year interval longitudinal studies and interpret the different results for MI-based bigrams.

Varieties of English for Specific Purposes dAtabase (VESPA) learner corpus (Paquot et al. 2013).[5] The corpus was part-of-speech (POS) tagged and parsed with the Stanford CoreNLP, and collocational strength was determined with MI on the basis of *amod*, *advmod*, and *dobj* Stanford-typed dependencies (cf. Section 2.2). Pairwise comparisons between groups revealed the following:

- Adjective + noun dependencies showed a significant difference in mean MI scores between the B2 and C2 levels, but differences were not large enough to distinguish between adjacent levels such as B2–C1 and C1–C2.
- Adverbial modifiers (i.e. adverb + adverb, adverb + adjective, adverb + verb) singled out upper intermediate (i.e. B2) learner writing from the more advanced (i.e. C1 and C2) learner productions.
- Verb + direct object structures were the best discriminators of the most advanced (C2) level: mean MI scores on *dobj* dependencies set C2 texts apart from B2 and C1 texts.
- No statistically significant difference was found between learner groups when their texts were analyzed with traditional measures of syntactic or lexical complexity.

In a follow-up study, Paquot (2018) made use of mixed-effects modeling to assess the influence of syntactic, lexical, and phraseological complexity on human raters' overall judgment of writing quality in the VESPA corpus. After stepwise model selection, the final model only included two mean-based phraseological measures, i.e. mean MI score for *dobj* dependencies and mean MI score for *amod* dependencies, as fixed effects (marginal R2 = 0.25), thus demonstrating that the higher the average MI scores for *dobj* and *amod* dependencies in a student's paper, the better it was assessed on the CEFR scale.

From the above, it can be argued that studies of syntactic co-occurrences in learner language can usefully complement the body of research based on a positional model of co-occurrence, most particularly by shedding light on how collocational strength of specific structures such as verb + direct object relations can be used to describe L2 performance and assess L2 proficiency. No study so far, however, has examined whether syntactic co-occurrences can also serve to trace phraseological development in a longitudinal learner corpus. Theoretically, Paquot (2019) has argued that L2 complexity

[5] https://uclouvain.be/en/research-institutes/ilc/cecl/vespa.html

research needs to broaden its scope on the ground that traditional measures of syntactic and lexical complexity fail to account for the fact that words naturally combine to form conventional patterns of meaning and use (cf. Sinclair 1991; Hanks 2013). This is particularly relevant since complexity is regarded as one of the "major variables in applied linguistic research" (Housen & Kuiken 2009): measures of linguistic complexity are widely used to describe L2 performance, assess L2 proficiency, and trace L2 development (Housen et al. 2012; Norris & Ortega 2009; Wolfe-Quintero et al. 1998). Following Ortega (2003, 492), the author offers the following working definition of phraseological complexity: "the range of phraseological units that surface in language production and the degree of sophistication of such phraseological units" (Paquot 2019, 4). Thus, a learner text with a wide range of (target-like) phraseological units and a high proportion of sophisticated units will be said to be more complex than one where the same few basic word combinations are often repeated.

The main objective of the present work is therefore to investigate phraseological complexity development in French EFL learner writing from the LONGDALE. The focus is placed on verb + object structures, as they have typically not been investigated in studies that adopted a structural model of statistical co-occurrence but, as mentioned above, are otherwise considered a major stumbling block for EFL learners. Building on the current state of the art, the study addresses the following research questions:

RQ1: To what extent can syntactic co-occurrences, and verb + direct object structures more particularly, be used to trace the development of phraseological complexity in a longitudinal corpus of EFL learner writing?

RQ2: What are the effects of proficiency vs. time spent learning English on phraseological complexity in learner writing development?

2 Data and Analysis

To answer the research questions, the study replicates the methods used to extract and analyze syntactic co-occurrences from the VESPA in Paquot (2018, 2019) on the LONGDALE corpus described in Section 2.1. Section 2.2 summarizes the different methodological steps required for that purpose and Section 2.3 reports the results of the statistical evaluation, which are discussed in Section 3. The final section closes with concluding remarks.

Table 24 Number of texts in LONGDALE sample used for this study

	Number of texts (with OQPT scores)
Y1 (2008 or 2010)	184
Y2 (2009 or 2011)	109
Y3 (2010 or 2012)	124
Total	**417**

2.1 Data

The learner data come from LONGDALE, a learner corpus compilation initiative that was launched in 2008 by the Centre for English Corpus Linguistics (UCLouvain) with the aim of collecting learner productions over a minimum period of three years, with data collections organized at least once a year (cf. Meunier 2016).[6] The subset used for this study consists of 417 argumentative essays written by two cohorts of French-speaking undergraduate students (total = 237) of English language and literature at UCLouvain, Belgium followed from 2008 to 2010 and from 2010 to 2012, respectively (Table 24).

The students wrote their essays in a computer lab using Notepad, with no access to reference tools. During the allotted time (90 minutes), students were requested to fill in a learner profile questionnaire, take a short vocabulary placement test (see below), and write an essay of about 500 words. In Year 1 (Y1), students had to choose one topic among the following: 'In our modern world, dominated by science, technology and industrialization, there is no longer a place for learning and imagination', 'Violent films are harmful and should be banned', 'Money is the root of all evil', and 'Lying is immoral and should always be condemned'. They could choose a topic in Y2 too (with four different prompts for Cohort 1, see Table 25) but were given the same topic in Y3 as the one they had selected in Y1 to ensure maximum comparability (Gentil & Meunier 2018, 276). As reported in Bestgen and Granger (2018, 282–283), "[a]ll the students proved to have enough time to write their essay. Although the exercise was compulsory, it was not part of a formal exam. However, the students did take it seriously, as they were promised individual feedback on the quality of their text."

[6] See also https://uclouvain.be/en/research-institutes/ilc/cecl/longdale.html.

Table 25 Prompts used in the LONGDALE sample

Code	Topic	Y1	Y2	Y3
modern	In our modern world, dominated by science, technology and industrialization, there is no longer a place for learning and imagination.	71	19	42
violence	Violent films are harmful and should be banned.	58	15	35
lying	Lying is immoral and should always be condemned.	28	16	22
money	Money is the root of all evil.	27	18	25
mothers	Mothers should stay home with their children.	0	15	0
media	The media pay too much attention to the personal lives of famous people.	0	13	0
judging	One should never judge a person by external appearances.	0	7	0
self	Self-confidence is the most important factor for success.	0	6	0

Each learner text comes with results from the Oxford Quick Placement Test (OQPT), one of several language tests used in the LONGDALE project to provide a measure of proficiency that is independent of learners' productions. The OQPT comprises 60 questions measuring knowledge of vocabulary and grammar; it has often been used as an indicator of general proficiency both in higher education (e.g. Meurant 2009)[7] and L2 research (see e.g. Hawkins et al. 2012).

Following Meunier and Littré (2013), OQPT scores were also converted into Common European Framework of Reference for Languages (CEFR) bands (Council of Europe 2001, 114), according to the following key: A1 (0–17), A2 (18–29), B1 (30–39), B2 (40–47), C1 (48–54), and C2 (55–60) for ease of interpretability. In this study, we used both OQPT scores and CEFR scores in separate models (cf. Section 2.3): we wanted to be able to compare our results with previous CEFR-based research while at the same time benefiting from the use of a potentially statistically more powerful numeric variable in the form of OQPT scores.

Crucially, learners at Y1 display a broad range of proficiency levels, from A2 to C2, with 79 percent at B1/B2, which makes it all the more important to investigate the role of proficiency vs. longitudinal development (RQ2) in LONGDALE.

[7] See also for example www.universiteitleiden.nl/en/language-centre/about-atc/information-for/prospective-students/english-language-assessment-for-ma-students.

Table 26 Number of EFL learners in LONGDALE sample

Trajectory	Number of students
Y1	86
Y1–Y2	17
Y1–Y2–Y3	66
Y1–Y3	15
Y2	10
Y2–Y3	16
Y3	27
Total	**237**

As discussed by Meunier and Littré (2013), attrition is one of the major challenges in dealing with longitudinal data, and the LONGDALE is no exception: Table 26 describes trajectories found in the learner corpus and shows, among other things, that 86 first-year students dropped out after the first year and 17 more did not provide data at Y3. Conversely, Table 26 also shows that 26 participants joined the study after the first year (Y1). Several methods have been used to deal with drop-in and drop-out phenomena in the LONGDALE. The first method is to use a sample of carefully selected learner texts, as for example did Bestgen and Granger (2018), who selected a subset of LONGDALE texts produced by the same learners at Year 1 and Year 3. The drawback of such an approach, however, is that corpus size can quickly drop dramatically. Consequently, we opted for mixed-effects modeling, as the technique makes it possible to explore the effects of different variables while dealing with unbalanced or missing data (cf. Field et al. 2012; Cunnings & Finlayson 2015; Gries 2015). Such an approach has been used, for example, by Meunier and Littré (2013) in their study of the development of accuracy in tense and aspect usage.

2.2 Data Preparation: Co-occurrence Extraction and Analysis

To investigate verb + direct object structures, we first made use of Ucto[8] and the TreeTagger (Schmid 1994) to tokenize, lemmatize, and POS tag each LONGDALE text; we then used the MaltParser (Nivre

[8] https://languagemachines.github.io/ucto/

et al. 2007) to parse all learner texts (engmalt.linear-1.7.mco model).[9] The next steps consisted in extracting from each LONGDALE text all the verb + noun pairs of words found in *dobj* Stanford-typed dependency relations in the form of lemmas and simplified POS tags. As illustrated in (1), a Stanford-typed dependency is a binary grammatical relation between a governor and a dependent (cf. de Marneffe & Manning 2013).

(1) *dobj* direct object
He won the lottery. *dobj*(win + VV, lottery + NN)

For each dependency, the total frequency of each individual word in each pair was recorded as well as their combined frequency.

Unlike in previous similar research (e.g. Durrant & Schmitt 2009; Granger & Bestgen 2014; Bestgen & Granger 2018), the study will only report on what MI scores have to reveal about EFL learners' use of statistical collocations, on the following grounds:

- T-scores have repeatedly been criticized for not having "a very transparent mathematical grounding" (Gablasova et al. 2017; cf. also Evert 2005, 82–83); they are also strongly dependent on corpus size (Gablasova et al. 2017, 169).
- In several studies, t-scores have been shown to be largely uninformative about EFL learners' use of syntactic collocations and regression modeling unable to explain much of its variance on the basis of fixed effects such as learner proficiency, topic, or time (e.g. Paquot & Naets 2015a, 2015b).
- By promoting the relatively less frequent and more semantically complex word pairs in learner productions, MI can be used as a measure of the 'sophistication' of word combinations (Paquot 2019).

Bestgen and Granger (2018) made use of the BNC, a 100 million-word collection of samples of written and spoken language from a wide range of sources designed to represent British English from the late twentieth century, to compute association measures and assign those values to word pairs extracted from LONGDALE texts. For the purposes of this study, however, we opted for the larger (9,578,828,861 tokens) and more recent Web corpus ENCOW14 (sentence shuffle AX version; Schäfer 2015).[10] As shown by Paquot and Naets (2017),

[9] www.maltparser.org/mco/english_parser/engmalt.html
[10] http://corporafromtheweb.org/encow14/

ENCOW14 will probably be a better choice when EFL learners' argumentative essays are investigated for at least the following two reasons:[11]

- ENCOW14 includes many co-occurrences that are perfectly idiomatic in English (e.g. consult + *dictionary*, *disprove* + *hypothesis*, *win* + *election*) but appear with too low frequency (fewer than five occurrences) in the BNC to be assigned an association score (e.g. Bestgen & Granger 2018, 284).
- Low frequencies, unreliable frequencies, or lack of appearance in the BNC can be attributed to too small a corpus size for collocation extraction and corpus age (Paquot & Naets 2017; cf. Brezina & Gablasova 2015 for a related discussion). If a larger and more representative corpus of today's English is used, the percentage of learners' co-occurrences used to compute indices of collocational strength also increases, thus improving the validity of such measures: focusing on verb + noun co-occurrences in the ICLE, Paquot and Naets (2017) showed that, on average, 33 percent of co-occurrences in each learner text were not used to compute a mean MI score when the BNC was used as a reference corpus; by contrast, only 6 percent were discarded when the ENCOW14 was used.

The ENCOW14 AX version also has the advantage of being distributed with Stanford-typed dependencies.

A list of *dobj* dependency-based pairs of words with frequency information was then used as input to the Ngram Statistics Package (NSP),[12] and an MI score was calculated for each word pair that appears with a frequency of at least five occurrences in ENCOW14.[13] Each word pair in the LONGDALE learner texts was

[11] In Paquot and Naets (2017), we also advocate the use of more than one reference corpus to assess learners' use of co-occurrences (see also Gablasova et al. 2017 for similar ideas).

[12] http://www.d.umn.edu/~tpederse/nsp.html

[13] Note that differences in proficiency level as represented in the LONGDALE learner texts have a small but not significant effect on the percentage of relational co-occurrences used to compute mean MI scores per text: a mean percentage of 87 percent of verb + object relations (tokens) are used to compute mean MI scores at A2, compared to 93 percent at C2. This is because the less proficient the learners, the more non-attested co-occurrences they use (from a mean of 2.7 unattested co-occurrences at B2 to 1.5 at C1). Spelling mistakes are also more frequent in the lower proficiency texts. In this study, we do not analyze these 'absent' co-occurrences (but see Bestgen & Granger 2014 for a study that makes use of this specific category to describe learners' collocational competence). In terms of frequency, mean MI scores are computed on a mean of 15.95 verb + object relations (sd = 4.99) at A2 (mean text length = 606, sd = 158), while they are computed on a mean of 18.56 verb + object relations (sd = 5.4) at C1 (mean text length = 633, sd = 89).

Table 27 Corpus preprocessing workflow

	Tools	Corpus
1. Tokenization	Ucto	LONGDALE
2. Lemmatization	TreeTagger	LONGDALE
3. POS tagging		
4. Parsing	MaltParser (engmalt. linear-1.7.mco model)	LONGDALE
5. Extraction of dependencies	In-house Perl programs	ENCOW14 +
6. Simplify POS tags		LONGDALE
7. Compute corpus-based frequencies		
8. Compute MI scores for each pair of words in a dependency	In-house Perl program (using Ngram Statistics Package)	ENCOW14
9. Assign MI scores computed on the basis of the ENCOW14 to each pair of words in a dependency in each learner text	In-house Perl program	LONGDALE
10. Compute mean MI scores for each learner text	R	LONGDALE

then looked up in the list of dependencies extracted from the ENCOW14 to determine its MI score in a reference corpus of contemporary general English. If a word pair was not found or appeared fewer than five times in ENCOW14, it was removed from further analysis. The last step involved computing a mean MI score for each learner text on the basis of all the different word pairs found in the *dobj* dependencies (i.e. types). Mean association measures were calculated with R (R Core Team 2014). Table 27 summarizes the different steps of the corpus pre-processing workflow; the workflow applied to LONGDALE texts aimed at replicating as accurately as possible the way ENCOW14 was processed by its compilers.

The analysis of relational co-occurrences requires accurate automatic syntactic analysis. To verify the quality of our dataset, we carried out a precision and recall study of *dobj* dependencies in 50 learner texts from the LONGDALE. While precision proved reasonably good (88.8 percent), the 76.9 percent recall rate obtained mainly stems from two major issues, i.e. POS tagging errors and erroneous dependency attachments (e.g. *people* is parsed as the subject of *judging* in 'many employers reject qualified people judging them on their appearance').

2.3 Statistical Evaluation

To answer our research questions, we made use of mixed-effects modeling to assess the influence of time (YEAR) on mean MI scores for *dobj* dependencies and investigate how it compares with the effect of proficiency, while taking into account any random variation observed across the participants (LEARNER). Proficiency effects were explored in terms of OQPT scores (OQPT) and CEFR levels (CEFR). We also included topic (TOPIC) as a fixed effect in our models because previous research has identified topic influence as an important explanatory factor for the presence or absence of specific word combinations in various text types and genres (e.g. Cortes 2004; Paquot 2014, 2017). All statistical analyses were performed with R (R Core Team 2014) and the *lme4*, *effects*, and *MuMIn* packages

We began our analysis with an exploration of the main variables involved in the study. These exploratory steps led to some minor changes that were required or at least useful for the subsequent regression modeling. Specifically, we studied all predictors and the response univariately and in a pairwise fashion both numerically (with summary statistics) and visually (with histograms) to check for outliers, data sparsity issues, potential curvature, etc. As a result:

- We had to conflate several levels of the predictor TOPIC to avoid massive data sparsity for all essay topics only written about by students in Y2. Therefore, the essays on 'judging', 'media', 'mothers', and 'self' were conflated into a group labeled *other year 2*. In the absence of specific a priori hypotheses about the differences between topics, we then ordered the factor levels by their descriptive mean MI scores and defined orthogonal contrasts for sequential-differences testing.
- We had to conflate levels of the predictor CEFR given the quite small frequencies of the levels A2 and C2, which were combined with B1 (to a factor level A2/B1) and C1 (to a factor level C1/C2), respectively. Given the ordinal nature of this predictor (from A2/B1 to B2 to C), CEFR was also set to utilize orthogonal sequential-differences contrasts.
- The variable YEAR was set to utilize orthogonal sequential-differences contrasts.
- The numeric predictor OQPT was centered to facilitate the interpretation of the coefficients in our regression models, i.e. from every value of OQPT we subtracted OQPT's overall mean of 43.1175.

In order to then determine the degrees to which the response variable MI is correlated with the predictors, we undertook a model selection and comparison approach using linear mixed-effects

modeling, which was conducted in several steps. First, we fitted a linear mixed-effects model with MI as the response variable, an overall intercept and the predictor TOPIC as fixed effects, and varying intercepts for participants; following Barr et al. (2013), we considered the possibility of a more comprehensive random-effects structure, but the nature of our data ruled that out: none of the speakers provided more than three data points (so the amount of repeated-measurements structure is minimal) and slightly more than 50 percent of the speakers provided more than one data point. (Some might argue that a mixed-effects model would not be required for these data, a point to which we will return briefly below.)

To determine which variable among YEAR, OQPT, and CEFR is the most useful predictor and answer RQ2, we followed the logic of Gries (2018); applied to this case, it means one needs to check which of these three predictors is most powerful and (i) whether Occam's razor then permits *adding* one or more of the others to the regression model already involving the most powerful predictor, or (ii) whether model comparison shows that in fact one or more of the other predictors (and potentially their interactions) can *replace* the most powerful predictor. Thus, we inspected the initial model involving only TOPIC and checked which predictor – OQPT, CEFR, or YEAR – would improve this model most (using both LR-tests and *AICc*-values, i.e. *p*-values using the traditional 0.05 threshold, and *AIC*-values corrected for smaller sample sizes). Both diagnostics showed that OQPT improved the model most, and OQPT was therefore added to the model. Following the above logic, we then tested both adding to and replacing OQPT. Model comparisons using *AICc* revealed that the model with TOPIC and OQPT (*AICc* = 792.27) was superior to:

- a model with YEAR instead of OQPT (*AICc* = 819.39);
- a model with CEFR instead of OQPT (*AICc* = 805.6);
- a model with YEAR and CEFR and their interaction instead of OQPT (*AICc* = 811.14).

In other words, the (preliminarily) final model contained only TOPIC and OQPT as predictors. However, an analysis of the coefficients for the predictor TOPIC showed that the five levels of the variable TOPIC were not justifiable from the perspective of Occam's razor; specifically, the only successive-differences contrasts that reached standard levels of significance ($p = 0.0454$) indicated a split between two kinds of topics: *money/modern/other* on the one hand vs. *violence/lying* on the other. We therefore created a binary version of TOPIC, i.e. one that only maintained these two levels and, to arrive at the most parsimonious model as well as determine that the results were not

Table 28 Summary results of the final regression model

	***b* / estimate**	***se***	***df***	$t_{\text{Satterthwaite}}$	***p***
Intercept	1.529	0.041	241.7	37.224	<<<0.001
$\text{TOPIC}_{\text{group1} \rightarrow \text{group2}}$	0.351	0.063	351	5.609	<<<0.001
$\text{OQPT}_{0 \rightarrow 1 \text{ etc.}}$	0.027	0.004	382.1	6.882	<<<0.001
Varying intercepts	PARTICIPANT	Residual			
	0.03	0.3			

affected in any way by the (lack of significant) differences between different topics, ran the above modeling process again.

As it turns out, the overall results were not affected by the simplification of the TOPIC variable: again, the final model involved just two predictors, TOPIC (now in its binary form) and OQPT (centered). The final statistics of this model are presented summarily in Table 28.

This model came with only a moderate amount of variance explanation, which, however, is unsurprising given the high degree of variability of such data and the small number of predictors involved: R^2_{marginal} (the *R*-squared value quantifying the summed effect of all fixed-effects predictors) is 0.16, whereas $R^2_{\text{conditional}}$ (the *R*-squared value quantifying the summed effect of both fixed and random effects) is 0.22. This indicates that the fixed effects account for more variability than the random ones, which is a positive sign – the data are not just mostly speaker-specific variable – but it also indicates that including the random-effects structure was useful. Model diagnostics, in particular of the residuals, revealed no problematic aspects of the final model and suggested no curved relationship between MI and OQPT; neither did a test of a polynomial to the second degree of OQPT.

The nature of the fixed effects is relatively straightforward and shown in the two figures below. Figure 3 shows the predicted mean MI scores on the *y*-axis as a function of the predictor TOPIC on the *x*-axis. The effect is indicated with points (the size of which is proportional to the number of data points per group) and their 95 percent confidence intervals; the grey points represent the actually observed values by the speakers. As is obvious from Table 28, the average predicted MI for the essays on *money*, *modern*, and *other* is 1.53 and significantly lower than the average predicted MI of 1.88 for the essays on *violence* and *lying*.

The effect of OQPT is shown in Figure 4: again, the MI scores are on the *y*-axis and the predictor is on the *x*-axis (which, for ease of interpretation, we de-centered). The regression line indicates the

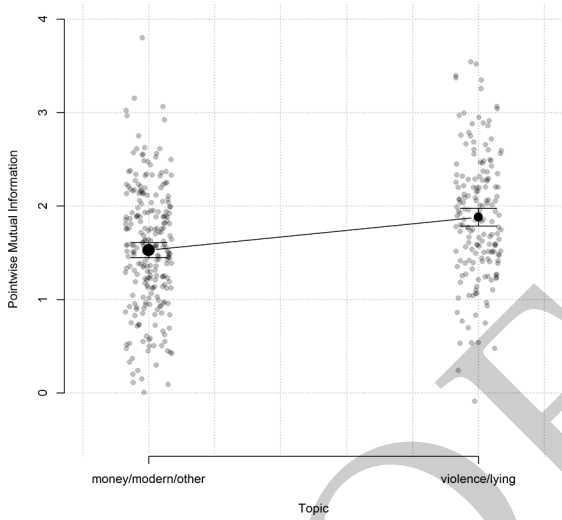

Figure 3 The main effect of TOPIC

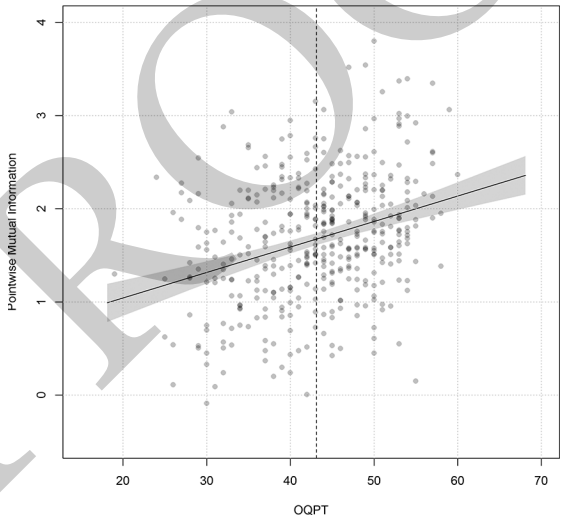

Figure 4 The main effect of OQPT

positive effect of OQPT on the mean MI scores (with a 95 percent confidence band) – the higher OQPT, the higher the mean MI scores – and the observed values are shown as grey points; the vertical line represents the mean of OQPT.

3 Discussion

Figure 5 represents the changes in average MI scores for *dobj* dependencies from Y1 to Y2 to Y3, with different colors representing different learner overall trajectories. As is clear, not all learners behave in the same way, and there is much variation in developmental trajectories at the individual level. As presented above, however, a mixed-model approach shows that, despite the apparent mess in data structure, there are also general patterns. First, the model reveals a significant effect of the prompt, with essays on 'Violent films are harmful and should be banned' and 'Lying is immoral and should always be condemned' featuring a higher MI mean value than the other essays. The effects are more subtle than just the typical reuse of word combinations primed by the prompt as documented in the literature (cf. for example, Ohlrogge 2009). For example, only 23.8 percent (436/1,835 tokens) of the *dobj* dependencies used in essays on

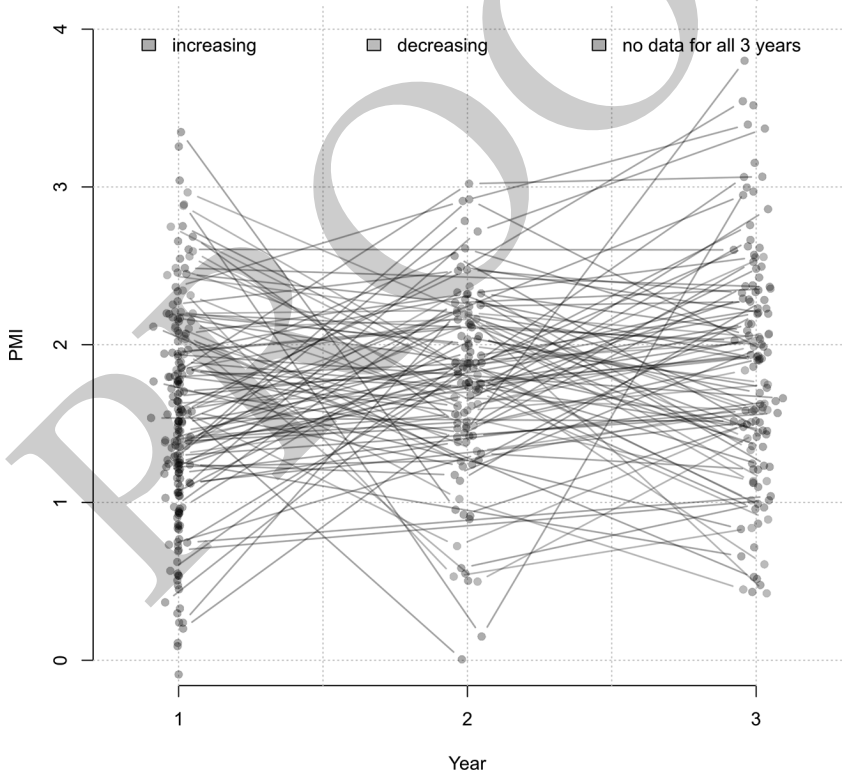

Figure 5 Average MI scores for dobj *dependencies per year*

'Money is the root of all evil' have a MI score above 3, which is a threshold often used in the literature for collocational status (e.g. Durrant & Schmitt 2009; Granger & Bestgen 2014). Examples of collocations that appear at least twice in the learner essays include *win + lottery*, *wear + clothes*, *play + role*, *cause + damage*, *pay + attention*, *break + rule*, *waste + money*, *spend + time*, *feed + family*, *donate + money*, *win + war*, *earn + money*, *reach + goal*, *spend + money*, *cause + problem*, *make + difference*, *bring + happiness*, *buy + house*, *imagine + world*, *buy + car,* and *make + profit*. Many other combinations of high-frequency words, including combinations with *money*, the theme of the prompt, have very low (if not negative) MI scores: *have + money*, *use + money*, *bring + money*, *receive + money*, *have + house*, *have + salary*, *buy + thing*, *have + value*, *give + food*, and *have + job*. By contrast, more than 30 percent (412/1,312 tokens) of the *dobj* dependencies used in EFL learner essays on 'Lying is immoral and should always be condemned' have a MI score above 3, with examples as varied as *mow + lawn*, *commit + crime*, *answer + question*, *serve + purpose*, *solve + problem*, *play + role*, *achieve + goal*, *cross + line*, *cause + damage*, *witness + murder*, *pay + attention*, *break + law*, *ask + question*, *hurt + feeling*, *keep + secret*, *break + vase*, *spend + time*, *cause + trouble*, *save + life*, *make + mistake*, *tell + story*, *reveal + truth*, *discover + truth*, *hide + truth*, *tell + truth*, *tell + lie*, *make + difference*, *change + world*, *organize + party*, *give + chance*, *deny + fact*, *protect + child*, and *give + example*. From the examples above, it seems that to answer the 'money' prompt, EFL learners only need to mobilize a limited set of related semantic domains that are made up of highly frequent words (frequently used nouns such as *money*, *food*, *family*, *job*, *time*, *war*, *house*, and *car*, and verbs such as *have*, *buy*, *use*, *give*, *win*, *play*, and *pay* belong to the 400 most frequent words in the *Corpus of Contemporary American English* (COCA).[14] By contrast, the 'lying' prompt mobilizes more semantic domains and less frequent words (e.g. *lawn*, *lie*, *truth*, *goal*, *murder*, *secret*, *trouble*, *commit*, *serve*, *solve*, *achieve*), which can be explained by the variety of lying examples and anecdotes that are found in such essays as exemplified below:

- *First, there are the ""kind"" lies. Those that cannot hurt people, like a man saying to his wife that he* ***mew*** *the* ***lawn*** *when he did not.* (UCL0211_Y1)
- *Those lies can be used out of laziness, for instance with a husband who did not* ***mow*** *the* ***lawn*** *and prune the bushes as he had*

[14] www.wordfrequency.info/free.asp?s=y

promised to his beloved wife while he was having a day off. (UCL0211_Y3)

- *On the other hand, some people lie because they have to escape from a situation, from a person. For instance, they* ***committed*** *a* ***crime*** *and do not want to be sent to jail.* (UCL0256_Y1)

To provide a first partial answer to RQ1, our results thus suggest that statistical co-occurrences, and verb + object relations more particularly, will only be useful to trace phraseological complexity development in longitudinal learner data if and only if topic/prompt is controlled or used as a predictor.

The second, and last, significant predictor in our final model is OQPT, i.e. a continuous numeric variable that represents scores on the OQPT. As shown in Figure 4, the higher the OQPT score of a learner, the higher their MI mean score. OQPT selection in the final model instead of the predictor YEAR answers RQ2: learner proficiency, as assessed by an independent measure (i.e. a standardized test), is a better predictor of phraseological complexity in each learner writing sample than the actual time when the essay was written (at the start of the curriculum, after one year or after two years of English instruction at university). This may be explained by two different factors:

(1) As shown in Figure 6, not all learners start with the same proficiency at Y1 (i.e. first year of a Bachelor program in English language and literature) as represented in the LONGDALE.[15] Quite the contrary, learners range from A2 to C2.
(2) The foreign language proficiency of each learner develops at a different pace. Table 29 illustrates the variety of individual trajectories in the LONGDALE.

Our results thus suggest that the time spent learning English will not have an effect on collocation strength per se. What matters more is foreign language proficiency (cf. Bestgen & Granger 2014) and whether learners improve from one year to the next: if the general foreign language proficiency of a learner does not improve from one data point collection time to the next, there is no reason to expect more phraseological complexity in their written productions.

That OQPT gets selected in the final model instead of the CEFR predictor is interesting but perhaps not particularly surprising. Since its publication in 2001, the CEFR has become the most widespread

[15] In Figure 6 and Table 29, we report CEFR levels instead of OQPT results for ease of interpretability.

Table 29 Language proficiency development: Individual trajectories in LONGDALE

	Year 1	**Year 2**	**Year 3**
UCL0005	C2	C1	NA
UCL0006	B2	C1	C1
UCL0008	B1	NA	B2
UCL0009	B2	NA	C1
UCL0012	NA	C1	C1
UCL0015	C1	C1	C2
UCL0021	B1	B2	B2

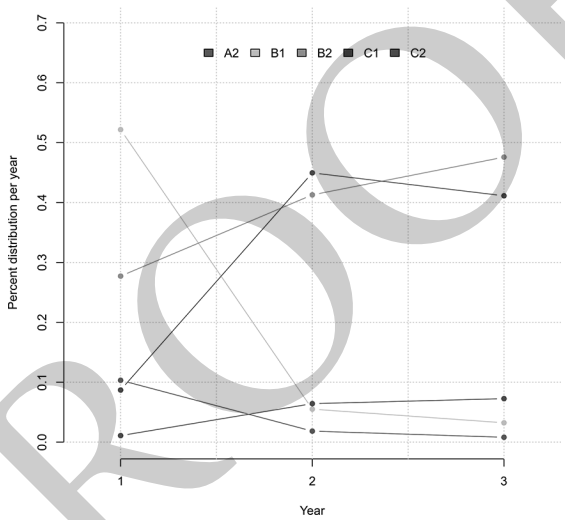

Figure 6 Percentage of A2 to C2 texts per year

reference tool in foreign language education and assessment across Europe, and learner corpus researchers have consequently seen a range of advantages in using a proficiency scale that is familiar to teachers, raters, and researchers in recent corpus compilation projects such as the KIAP or the MERLIN corpora (Carlsen 2012; Abel et al. 2014) as well as in post-hoc assessment of samples of learner texts from well-established learner corpora such as the ICLE (e.g. Thewissen 2013).[16]

[16] See Wisniewski (2017), however, for a discussion on why learner corpora should be linked only very carefully to the CEFR levels.

The use of (well-defined) proficiency categories also offers many advantages in Second Language Research (SLR), including ease of interpretation (i.e. learner groups at different proficiency levels can easily be compared) and enhanced comparability across studies, which in turn should ideally lead to more generalizability of the results and more practical outcomes. However, from a methodological/statistical perspective, our final model shows that the numeric predictor OQPT is more informative than its derived categorical predictor CEFR: it explains more variance in EFL learners' use of statistically assessed *dobj* dependencies in the LONGDALE. Our results thus support Ortega's call for more SLR study designs where proficiency is treated as an interval scale (i.e. individual proficiency scores), and not as a categorical variable as has most often been done in the field:

> SLA researchers have most often chosen to treat proficiency as a categorical variable and then have assessed mean differences in complexity values across proficiency groupings. Yet, this practice of converting interval variables (i.e. individual proficiency scores of some kind) into categorical ones (i.e. participants grouped by nominal proficiency levels) has always been criticized by statisticians because it discards much useful information. More specifically, it does away with the variance of continuous scores and leads to unreliability and increased likelihood of Type II errors (e.g. Skidmore and Thomson 2010), that is, the problem of failing to detect a difference, relationship, or effect that is in fact present because of some psychometric methodological problem, such as lack of power or (in the case at hand) lack of variance in the observations.
>
> (Ortega 2012, 131)

4 Conclusion

At EuroSLA2015, the first two authors reported on a monofactorial study that investigated the effect of time on mean MI scores in the LONGDALE; they used a mixed-effects modeling approach and showed that, on average, mean MI score per learner text increased significantly at a rate of 0.14 for every one unit increment in time (i.e. every year) (Paquot & Naets 2015b). However, the model performed badly, with the fixed effect explaining only 3 percent of the variance, and the random effects accounting for an additional 14 percent, i.e. between-speaker variability accounting for nearly five times as much as the fixed effect. In the new model presented in this study, the explained fixed-effects variance is five times as high (0.16) as in Paquot and Naets's (2015b) model, and what the between-speaker variability adds is only 7 percent. In other words, the fixed effects do five times as much as in the old model (which is good), while the

random effects do only half as much as in the old model (which is also good). To obtain such better results, we re-examined the same dataset as used in Paquot and Naets's (2015b) study, this time, however, adopting a multifactorial design that includes two predictors that have already been shown to have an effect on EFL learners' use of word combinations of various types, i.e. proficiency level and essay prompt/topic. By doing so, we followed Gries (2018)'s recommendation to approach even a monofactorial hypothesis (here, time has an effect on phraseological complexity development in EFL learner writing) with a multifactorial design "to determine either (a) whether it adds anything to what we already know about the phenomenon (by statistically controlling for what we already know) or (b) whether it replaces (parts of) what we already know about the phenomenon" (Gries 2018, 296). We were able to determine that the time dimension does not add anything to or modify what we already know about EFL learners' use of statistical co-occurrences (as represented in the LONGDALE). Thus, the significant effect of time reported by Paquot and Naets (2015b) needs to be taken with a grain of salt, given how an overall better degree of variance explanation is in fact obtained by the (correlated, but more fine-grained) predictor of proficiency as well as that of topic/prompt.

Our findings have other important implications for LCR in general and longitudinal studies more particularly. First, they call for a more systematic control of topic/prompt in phraseological studies of learner language samples than has been done so far, including in the authors' own work. Second, they point to the need to account (statistically) for individual variation and individual trajectories in longitudinal corpora: not only do learners not start with the same initial proficiency level in English in the LONGDALE (as they would also be expected not do in other longitudinal learner corpora)[17] but their foreign language proficiency also develops in different ways. This also means that foreign language development over time in the LONGDALE should ideally be investigated in tandem with the development of proficiency as measured by means of the OQPT. The availability of independent proficiency scores in LONGDALE is certainly an invaluable strength of the longitudinal corpus.

[17] The diversity of proficiency levels represented in Year 1 in the LONGDALE also provides empirical support to repeated calls that we (this is an inclusive 'we'!) should exercise more caution in the use and analysis of learner language samples for which the foreign language proficiency of learners is operationalized by institutional status (cf. Thomas 2006; Tono 2003; Carlsen 2012).

The study also comes with limitations that we would like to address in future research. First, although we improved the study design as compared with Paquot and Naets (2015b) and examined the effects of four predictors (topic, time, and proficiency as operationalized by CEFR and OQPT) on the development of phraseological complexity in EFL learner writing, the explained variance is still limited. We would like to investigate whether other learner variables (e.g. time spent in an English-speaking environment) and text/linguistic variables (e.g. lexical diversity, lexical dispersion, lexical sophistication) could increase the amount of explained variance in our dataset. Examining the potential effects of traditional measures of lexical and syntactic complexity, for example, would also make it possible to explore how the various dimensions of linguistic complexity interact with time and/or proficiency (cf. Paquot 2019). Second, we would like to use other association measures than MI, more particularly Delta P or the log odds ratio to explore collocation strength and phraseological complexity in learner language: both keep frequency and effect size separate, and the first is also directional. Last but not least, we are very much aware that a mean MI is a very crude measure of a learner's phraseological competence, and we are currently investigating how to use more information from each learner text.

Acknowledgments

We would like to thank Professor F. Meunier (Centre for English Corpus Linguistics, UCLouvain) for granting us access to a set of French learners' essays from the LONGDALE.

References

Abel, A., Nicolas, L., Wisniewski, K., Boyd, A., & Hana, J. (2014). A trilingual learner corpus illustrating European reference levels. *Ricognizioni. Rivista di Lingue e Letterature e Culture Moderne* 2(1), 111–126, https://doi.org/10.13135/2384-8987/702

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3), 255–278.

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

Bestgen, Y. & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing* 26, 28–41.

(2018). Tracking L2 writers' phraseological development using collgrams: Evidence from a longitudinal EFL corpus. In S. Hoffmann, A. Sand,

S. Arndt-Lappe, & L. M. Dillman (eds). *Corpora and Lexis*, 277–301. Leiden & Boston: Brill Rodopi.

Brezina, V. & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics* 36(1), 1–22.

Carlsen, C. (2012). Proficiency level – a fuzzy variable in computer learner corpora. *Applied Linguistics* 33(2), 161–183.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Cunnings, I. & Finlayson, I. (2015). Mixed effects modelling and longitudinal data analysis. In L. Plonsky (ed.) *Advancing Quantitative Methods in Second Language Research*, 159–181. Abingdon: Routledge.

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

De Marneffe, M.-C. & Manning, C. (2013). Stanford Typed Dependencies Manual, retrieved from http://nlp.stanford.edu/software/dependencies_manual.pdf (accessed June 13, 2020).

Durrant, P. & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL – International Review of Applied Linguistics in Language Teaching* 47(2), 157–177.

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-Based Approaches to Language Acquisition and Processing: Cognitive and Corpus Investigations of Construction Grammar*. Language Learning Monograph Series. Chichester: Wiley-Blackwell.

Evert, S. (2005). The Statistics of Word Cooccurrences: Word Pairs and Collocations. Unpublished Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, retrieved from www.collocations.de/phd.html (accessed June 13, 2020).

(2008). Corpora and collocations. In A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics: An International Handbook*, 1212–1248. Berlin: Mouton de Gruyter.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics with R*. Los Angeles, CA: SAGE Publications Ltd.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning* 67(S1), 155–179.

Gentil, G. & Meunier, F. (2018). A systemic functional linguistic approach to usage-based research and instruction. The case of nominalization in L2 academic writing. In A. E. Tyler, L. Ortega, M. Uno, & H. I. Park (eds.), *Usage-Inspired L2 Instruction. Researched Pedagogy*, 267–289. Amsterdam and Philadelphia: Benjamins.

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.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *The International Corpus of Learner English, Version 2*. Handbook and CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S. & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (eds), *Phraseology: An Interdisciplinary Perspective*, 27–49. Amsterdam & Philadelphia: Benjamins.

Gries, S. Th. (2015). The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10(1), 95–125.

(2018). On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies* 1 (2), 277–309.

Gries, S. Th. & Wulff, S. (2005). Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics* 3, 182–200.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press.

Hawkins, R., Althobaiti, M., & Ma, Y. (2012). Eliminating grammatical function assignment from hierarchical models of speech production: Evidence from the conceptual accessibility of referents. *Applied Psycholinguistics* 35(4), 1–31.

Housen A. & Kuiken, F. (2009) Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics* 30(4), 461–73.

Housen A., Kuiken F., & Vedder, I. (2012) (eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Amsterdam: Benjamins.

Jones, S. & Sinclair, J. McH. (1974). English lexical collocations. *Cahiers De Lexicologie* 24, 15–61.

Kyle, K. & Crossley, S. A. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing* 34(4), 513–535.

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

Meunier, F. (2016). Introduction to the LONGDALE project. In E. Castello, K. Ackerley, & F. Coccetta (eds.), *Studies in Learner Corpus Linguistics: Research and Applications for Foreign Language Teaching and Assessment*, 123–126. Bern: Peter Lang.

Meunier, F. & Littré, D. (2013). Tracking learners' progress. Adopting a dual 'corpus cum experimental data' approach. *The Modern Language Journal* 97(1), 61–76.

Meurant, R. (2009). Computer-based Internet-hosted assessment of L2 literacy: Computerizing and administering of the Oxford Quick Placement Test in ExamView and Moodle. In D. Ślęzak, W. I. Grosky, N. Pissinou, T. K. Shih, T. Kim, & B. H. Kang (eds.), *Multimedia, Computer Graphics and Broadcasting. MulGraB 2009. Communications in Computer and Information Science, Vol. 60*, 84–91. Berlin & Heidelberg: Springer.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Amsterdam & Philadelphia: John Benjamins.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., & Marsi, E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13(2), 95–135.

Norris, J. M. & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics* 30(4), 555–578.

Ohlrogge, A. (2009). Formulaic expressions in intermediate EFL writing assessment. In R. Corrigan, A. Moravcsik, H. Ouali, & K. M. Wheatley (eds.), *Formulaic Language: Vol. 2. Acquisition, Loss, Psychological Reality, and Functional Explanations*, 387–404. Amsterdam: John Benjamins.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics* 24(4), 492–518.

(2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann & B. Szmrecsanyi (eds.), *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, 127–155. Berlin & Boston: Mouton de Gruyter.

Paquot, M. (2014). Cross-linguistic influence and formulaic language: Recurrent word sequences in French learner writing. In L. Roberts, I. Vedder, & J. Hulstijn (eds.), *EuroSLA Yearbook*, 216–237. Amsterdam and Philadelphia, PA: John Benjamins.

(2017). L1 frequency in foreign language acquisition: Recurrent word combinations in French and Spanish EFL learner writing. *Second Language Research* 33(1), 13–32.

(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(1), 29–43. DOI: https://doi.org/10.1080/15434303.2017.1405421 10.1080/15434303.2017.1405421

(2019). The phraseological dimension in interlanguage complexity research. *Second Language Research* 35(1), 121–145. https://doi.org/10.1177/0267658317694221

Paquot, M., Hasselgård, H., & Oksefjell Ebeling, S. (2013). Writer/reader visibility in learner writing across genres: A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora. In S. Granger, G. Gilquin, & F. Meunier (eds.), *Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead. Corpora and Language in Use – Proceedings 1*, 377–387. Louvain-la-Neuve: Presses universitaires de Louvain.

Paquot, M. & Naets, H. (2015a). Adopting a relational model of co-occurrences to trace phraseological development. Paper presented at the 3rd Learner Corpus Research Conference. Netherlands, September 11–13 2015.

(2015b). Using relational co-occurrences to trace phraseological development in a longitudinal corpus. Paper presented at the 25th EuroSLA conference. Aix-en-Provence, August 27–29, 2015.

(2017). The role of the reference corpus in studies of EFL learners' use of statistical collocations. Paper presented at ICAME38. Prague, May 24–28, 2017.

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, retrieved from www.R-project.org (accessed June 13, 2020).

Schäfer, R. (2015). Processing and querying large Web corpora with the COW14 architecture. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, & W. Andreas (eds.), *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)*, 28–34. Mannheim: Institut für Deutsche Sprache.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. Paper presented at the International Conference on New Methods in Language Processing. Manchester, UK.

Sinclair, J.M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97(S1), 77–101.

Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In J. Norris and L. Ortega (eds.), *Synthesizing Research on Language Learning and Teaching*. Amsterdam: John Benjamins.

Tono, Y. (2003). Learner corpora: Design, development and applications. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, & S. Khoja (eds.), *Proceedings of the Corpus Linguistics 2001 Conference*. UCREL Technical Paper Number 13, 800–809. Lancaster, Lancaster University.

Wisniewski, K. (2017). Empirical learner language and the levels of the Common European Framework of Reference. *Language Learning* 67(S1), 232–253.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. Honolulu, HI: University of Hawaii Press.