Published in *System 69* (2017) 65-78 http://dx.doi.org/10.1016/j.system.2017.08.004

Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence

Yves Bestgen

Abstract

Formulaic competence, the native-like use of ready-made sequences of words, is a key aspect in the development of L2 writing proficiency. Becoming increasingly important in foreign language teaching, it is largely neglected when assessing learner writing. Recently, several (semi-)automatic methods have been proposed to analyse the formulaic sequences in learner texts. After describing these methods and discussing their strengths and limitations, this study aims at determining the usefulness for assessing L2 text quality of collgram, a technique that assigns to each pair of contiguous words in a learner text two association scores (mutual information and t-score) computed on the basis of a large native speaker reference corpus. Correlation and hierarchical regression analyses, conducted on two datasets of English-as-a-foreign-language texts, showed that formulaic measures were the best predictors of text quality and provided a much higher specific contribution to the prediction than single-word lexical measures of diversity and sophistication. This study also confirms the need, as is being increasingly emphasized in applied linguistics, to replicate analyses on several corpora. The conclusion addresses the limits of the study and points to some potential implications for the teaching and assessment of second language writing.

Keywords: Formulaic sequence; L2 writing assessment; Lexical diversity; Lexical sophistication; Association measure; Mutual information; Replication study.

1. Introduction

The recognition of the importance of the formulaic dimension of language (Cowie, 1981; Pawley & Syder, 1983; Sinclair, 1991; Wood, 2015; Wray, 2012) has led to a large increase in studies focusing on various aspects of L2 formulaic language. Many of these studies have compared the use of formulaic sequences (FS), such as *brief overview*, *depend on, out of, by the way,* and *as far as I know*, by native and non-native speakers. The general conclusion is "that learners tend to have a small inventory of formulaic sequences that they overuse" (Wray, 2012, p. 235). Only very advanced learners can be expected to show knowledge of FSs that is similar to that of native speakers (Boers & Lindstromberg, 2012). Cross-sectional studies, which compare learners at different proficiency levels, and longitudinal studies have confirmed this genuine, but slow, development of formulaic competence when learning a foreign language (Appel & Wood, 2016; Huang, 2015; Li & Schmitt, 2009; Qi & Ding, 2011; Siyanova-Chanturia, 2015; Verspoor, Schmid, & Xu, 2012).

It is therefore not surprising that FSs are receiving increasing attention in foreign language teaching. Recently, several lists of the most pedagogically useful FSs, especially in academic discourse (Ackermann & Chen, 2013; Byrd & Coxhead, 2010; Simpson-Vlach & Ellis, 2010; Wood & Appel, 2014), were compiled, extending well-known lists of pedagogically important words (Coxhead, 2000). Sophisticated methods have also been proposed to analyse the FSs in texts either automatically or semi-automatically (Crossley, Cai, & McNamara, 2012; Granger & Bestgen, 2014; Leńko-Szymańska, 2014; Staples, Egbert, Biber, & McClair, 2013; Treffers-Daller, Parslow, & Williams, 2012). Two of these techniques are now freely available on the Internet (Kyle & Crossley, 2015; Leńko-Szymańska & Wolk, 2016). It seems therefore that the field of formulaic language research is catching up with that of single-word lexical research, for which automated lexical

profiling tools have been available for a while (Cobb, 2013). This evolution raises the question of the effectiveness of formulaic measures for assessing L2 texts when compared to simpler and better-established single-word lexical measures. Trying to answer this question is the main objective of the study. It is hoped that this research will shed some light on how to measure the development of the formulaic competence, an important issue in second language learning (Huang, 2015), and will eventually make available to foreign language teachers effective tools for assessing the use of formulaic sequences (FS) by learners.

2. Literature review

2.1. Using formulaic measures in L2 text assessment

Analysing the use of FSs in learner texts is a much more complex task than analysing single words (Read, 2000). Different approaches are possible, from the most manual to the most automated. The manual approach establishes the formulaic nature of a lexical sequence by using dictionaries or a native speaker's intuition guided by a list of criteria (Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006; Verspoor et al., 2012). This approach is complex and costly, but allows for a very fine qualitative analysis. It is difficult to use for assessing L2 writing on a large scale, since manual identification of the FSs must be redone for each new set of texts.

The manual analysis of each new learner text can be avoided by using the frequency-based approach that makes use of the frequencies in a corpus to separate the FSs from the free combinations (Gyllstad, 2007). In learner corpus research, the most often used frequency-based approach is based on the automatic identification of recurrent continuous sequences of three or more words in a corpus (Biber, Johansson, Leech, Conrad, & Finegan, 1999). These *lexical bundles* are extracted on the basis of a frequency criterion (10 to 40 occurrences per million words) and a dispersion criterion (occurring in at least three texts). They are then classified according to their structural characteristics (e.g., noun phrase and prepositional phrase) and discourse functions (e.g., referential bundles and stance bundles). This approach has yielded many important results. Compared to native speakers, L2 learners have a more limited repertoire of lexical bundles that they tend to repeat often, such as *on the other hand* and *at the same time* (Ädel & Erman, 2012; Chen & Baker, 2010). Staples et al. (2013) showed that lower proficiency learners use more lexical bundles than higher proficiency learners, but that many of these bundles were copied from the prompts or the texts provided to them. The functional categorization of the lexical bundles has also been shown to be very useful in distinguishing proficiency level. Appel and Wood (2016) observed that, compared to lower proficiency learners, higher proficiency learners used more referential word combinations that identify an entity or define some of its attributes (e.g., *certain forms of aquatic life*). Looking at the formal/informal dimension in expository and argumentative essays, Chen and Baker (2016) found that lower-level learners tended to use more lexical bundles which typically occur in conversation, such as *there are a lot of* and *there are too many*, while the bundles used by higher-level learners were closer to those found in academic prose such as *a great deal of*. These promising results led several authors to recommend the use of lexical bundles to assess the quality of texts (Appel & Wood, 2016; Staples et al., 2013). As the differences found between lower-level and higher-level learners result from a manual analysis of the lexical bundles, additional work is necessary to automate this step to the greatest possible extent in order to allow the use of this approach in educational settings.

Several approaches have been proposed to try to avoid the use of human judgment. The most extreme one is probably that of Treffers-Daller et al. (2012, 2013), who use the frequency of all n-grams (i.e., any sequence of contiguous words) of different lengths in learner texts without even trying to determine which ones are formulaic and which ones are not. However, these n-grams based measures were much less effective than single-

word lexical measures to discriminate texts written by learners at different levels of the Pearson Test of English Academic. Moreover, these measures are not well suited for classroom use by foreign language teachers.

A more promising approach uses a large native speaker reference corpus as a proxy for native speaker intuition. Each potential FS present in the learner text is sought in this reference corpus in an attempt to determine if it is typical of native language use. This approach is well illustrated in a study by Leńko-Szymańska (2014) on the use of lexical bundles by L2 learners from six L1 backgrounds and three different academic grades. The three-word bundles were not retrieved directly from the learner texts as it was the case in the lexical bundles studies discussed above, but through a reference list made up of the most recurrent three-word sequences in the 425-million-word Corpus of Contemporary American English (COCA). This list was then used to identify the lexical bundles that occur in a learner text. She observed statistically significant differences in the number of lexical bundles between L1 groups and between grades. Qin (2014) used the same procedure to compare the use of lexical bundles by non-native English graduate writers and by published authors in the field of applied linguistics. She first identified the five-word lexical bundles that occurred in a one-million-word corpus of published articles in the domain, such as *at the end of the* and *English as a foreign language*, before searching for their use in learner and expert texts. She observed an increase in the diversity of lexical bundles used in relation to the year of study. As these two studies compared the lexical bundle frequencies between corpora, disregarding the individual texts, it remains to be shown that the approach can be used to reliably assign formulaic scores to individual texts.

Kyle and Crossley's (2015) technique combines those of Treffers-Daller et al. (2012) by taking into account all of the bigrams and trigrams in a learner text and those of Leńko-Szymańska (2014) and Qin (2014) by assigning them several scores based on how often they occurred in a large native speaker reference corpus (i.e., the British National Corpus). Analysing two English learner corpora, one of written texts and one of short spoken responses, Kyle and Crossley (2015) showed these n-gram scores are useful predictors of text quality. More frequent n-grams and a smaller proportion of n-grams absent from the reference corpus were associated with better lexical and speaking proficiencies. Kyle and Crossley have provided a calculator for these indices on a freely accessible web page (http://www.kristopherkyle.com/taales.html).

The procedure of Kyle and Crossley (2015) is entirely based on the frequency of n-grams in a reference corpus. Frequency is undoubtedly an important indicator for defining and identifying FSs, but, particularly in the case of bigrams, it has the disadvantage of not taking into account the frequency of the words that compose the sequence. A sequence can be highly frequent, not because of its formulaic status, but because it is made up of very frequently used words (Bestgen, in press; Evert, 2009). To overcome this weakness, a large number of lexical association measures have been proposed (Pecina, 2010). These are measures of attraction between words which usually compare the frequency of occurrence of a pair of words with the frequency that would be expected if the words were randomly ordered in the corpus. The higher the score, the more associated the words. These association measures have proved particularly effective in identifying collocations, which are a type of formulaic expression made of strongly associated pairs of words characterized by restricted substitutability (e.g., *powerful computer*, *target audience*, *strongly agree* and *highly controversial*). Durrant and Schmitt (2009; see also Siyanova-Chanturia, 2015) have chosen to use two of these measures, calculated on the basis of a large native speaker reference corpus, for analysing FSs in learner texts. They selected mutual information (MI) and t-score because each of these measures highlights a different type of expression. MI is a measure of strength of association, equal to the log-transformed ratio between the observed frequency of the bigram and its expected frequency. To get a large MI, a bigram has to be made up of words that are (somewhat) rarely found away from each other. It is

easier to fulfil this condition with bigrams made of rare words, such as *immortal souls*, *cayenne pepper* and *somewhat ambiguous*, than of frequent words (Clear, 1993). The t-score expresses the confidence about the existence of an association between two words. It is calculated by dividing the difference between the observed and expected frequencies of the bigram by the square root of its observed frequency. It prioritizes frequently occurring bigrams, such as *more than*, *you know* and *good example,* that must therefore be made up of frequent words (Clear, 1993).

Durrant and Schmitt's (2009) study was focused on bigrams, made up of a modifier (adjective or noun) followed by a noun, which were manually extracted from texts. Compared to native speakers, L2 learners tended to underuse lower-frequency, but strongly associated, collocations (identified by their MI score), while overusing high-frequency collocations (identified on the basis of t-score). More recently, Granger and Bestgen (2014) observed similar differences between intermediate and advanced L2 learners. Bestgen and Granger (2014) showed that the average MI score and the proportion of bigrams in the text that are absent from the reference corpus were correlated with text quality rating. The methodology used in these studies differed from Durrant and Schmitt's procedure in the three following ways: (1) the bigrams were extracted automatically from the texts; (2) the full range of bigrams present in the texts was investigated, and (3) all of the bigrams were set on a continuum from the most formulaic expressions (i.e. those with the highest association scores, such as *vicious circle* and *absolutely imperative* for MI and *great deal* and *too late* for the t-score) to the least (i.e. those bigrams that cooccur in the corpus less often than expected by chance, such as *include of* and *all person* for MI and *the all* and *of they* for the t-score). All of these properties ensure that these measures are readily usable for the purpose of automatically analysing learner texts. Because the formulaic units analysed combine the strengths of both collocations (by using association scores) and n-grams (by using contiguous pairs of words), Bestgen and Granger (2014) referred to them as *collgrams*.

Recently, the collgram procedure has been made freely available on the Internet by Leńko-Szymańska and Wolk (http://collgram.pja.edu.pl/) to automatically score a text and to show the collgrams present in it and their association scores. It is therefore worthwhile to confirm its effectiveness in estimating text quality, especially since using a reference corpus to identify FSs does not come without problem. Compared to native speaker intuition, it is obviously much coarser. Compared to the lexical bundle approach, it will lead to consider as non-formulaic sequences, such as *make an experience* and *achieve tasks* (Nesselhauf, 2005), that are used by learners and are formulaic for them (Wood, 2015), but that are rarely used by native speakers (Durrant & Schmitt, 2009). According to Oppenheim (2000), these idiosyncratic FSs made up almost all of the FSs occurring in learner speeches, at least when they were defined as a word sequence repeated verbatim when a speaker gave the same speech twice. It must also be noted that MI and t-score are only two of the association measures proposed in the literature (Pecina, 2010). They were selected because of their well-established complementarities for identifying FSs (Church, Gale, Hanks, & Hindle, 1991; Evert, 2005)

2.2. Formulaic and lexical richness measures of L2 competence

Many studies have shown that measures of the degree to which a varied and large single-word vocabulary is used in a learner text - its lexical richness as Laufer and Nation (1995) call it - are very good predictors of text quality (Crossley, Cobb, & McNamara, 2013; Crossley, Salsbury, & McNamara, 2012; Engber, 1995; Jarvis & Daller, 2013; Treffers-Daller et al., 2012, 2016; Yu, 2010). The main thesis of the present study is that analysing the FSs should also be useful for assessing the quality of learner texts because they are "a necessary component of second-language (L2) lexical competence in addition to the knowledge of single words" (Laufer & Waldman, 2011, p. 648). However, their

(supposed) effectiveness could result from the fact that evaluating the use of FSs is partly also evaluating the use of the single words that compose them. It is thus necessary to show that a formulaic measure makes a specific contribution to the prediction of the quality of a text when compared to more classical single-word lexical measures.

The two features of lexical richness that have been the most widely studied in foreign language writings (Daller & Xue, 2009) deserve special attention: *lexical diversit*y (Tweedie & Baayen, 1998), which is based on the number of different words in a text, and *lexical sophistication* (Laufer & Nation, 1995), which is based on the use of advanced words. They are two of the four lexical richness features, along with lexical density and the number of lexical errors, discussed by Read (2000) in his book on vocabulary assessment. Lexical density, which is the proportion of lexical (as opposed to grammatical) words in a text, is not analysed here because it is not considered to be particularly relevant for the assessment of writing (Malvern, Richards, Chipere, & Durán, 2004; Read, 2000) and because several studies did not show it to be related to text quality (Lu, 2012; Šišková, 2012; Vidakovic & Barker, 2010). Regarding lexical errors, Read (2000) emphasized that it is difficult to define exactly what must be counted as a lexical error and that it is necessary to take into account the seriousness of the different types of errors, an issue that goes far beyond the scope of this study. Lexical richness will later be used to subsume the single-word lexical measures of diversity and sophistication and it therefore will not refer to the formulaic measures, even if the use of FSs is obviously part of it.

A few studies have attempted to assess the strength of the relationship between formulaic and lexical richness measures in learner texts. Levitzky-Aviad and Laufer (2013) found moderate, yet statistically significant, correlations ($.20 < r\text{s} < .33$, $p < .01$, $N = 290$) between the use of manually identified verb-noun and adjective-noun collocations and three measures of lexical diversity and sophistication in texts written by learners of English. Forsberg Lundell and Lindqvist (2012) observed a correlation of the same intensity between the use of FSs and lexical sophistication among learners of French as a foreign language, but this correlation failed to reach statistical significance ($p = .06$), probably because of the small number of texts analysed ($N = 30$). Vedder and Benigno (2016), when analysing texts written by 39 lower-intermediate and intermediate learners of L2 Italian, did not observe any relationship between L2 language proficiency and either collocational competence or lexical richness. However, this study was based on a much narrower proficiency range than Levitzky-Aviad and Laufer's (2013) study, which examined evolution over eight years of learning. Analysing experts' ratings of learner texts, Crossley, Salsbury, and McNamara (2015) observed that holistic scores of lexical proficiency were more correlated to the assessment of a learner's accuracy in producing FSs than to several single-word lexical ratings. They also underlined the high correlation between single-word and multi-word ratings.

The limited number of studies that have addressed the relationship between formulaic and lexical richness measures preclude any kind of conclusion, especially since the intensity of this relationship should depend on the type of formulaic measures used. As explained above, MI shows a very strong frequency bias: it favours expressions composed of rare words, indicating a high level of lexical sophistication, as well as an increased lexical diversity because these words are seldom repeated. A formulaic measure based on MI could, therefore, be much more related to single-word lexical measures than those assessed so far.

3. Research questions

This study aims to answer the following two research questions:

- To what extent do the formulaic measures automatically extracted by the collgram procedure estimate the quality of learner texts?

- How important and how specific is their contribution to the prediction of text quality in comparison to single-word measures of lexical diversity and sophistication?

To answer these questions, formulaic and single-word lexical measures were applied to two datasets, which contained a large number of English-as-a-foreign-language texts that had been holistically evaluated, and their effectiveness in predicting text quality scores was compared. These analyses were conducted on two very different sets of learner texts because replicating analyses on different corpora is becoming increasingly important in applied linguistics to strengthen the conclusion (Porte, 2012).

4. Material and methods

4.1. Datasets

Two English learner datasets were analysed in order to increase the degree of generality of the conclusions. These datasets were selected partly because any spelling errors they contained had been manually annotated, which is a prerequisite for the proper application of lexical richness measures (Kojima & Yamashita, 2014; Laufer & Nation, 1995).

4.1.1. The FCE dataset

The first dataset was composed of the First Certificate in English (FCE) examination scripts (Yannakoudakis, Briscoe, & Medlock, 2011a, 2011b). This test, which forms part of the Cambridge Assessment's English as a Second or Other Language examinations, was designed to evaluate upper-intermediate learners of English. Extracted from the Cambridge Learner Corpus, the dataset consisted of 1,235 documents of between 200 and 400 words taken from the written section of the examination, i.e., a total of 460,964 words. Each document included two texts[1] written by the same learner in response to prompts, the first being a mandatory letter and the second a free choice between another letter or a composition, a story, a report or a magazine article. Typical prompts used for collecting these texts are provided in Appendix 1. The participants had 80 minutes to write both texts. They came from 16 different L1 backgrounds, the most frequent of these being Spanish (16%) and French (11.7%), but there were also at least 5% each of German, Chinese, Japanese, Korean, Thai, Turkish and Polish participants. While the majority of the test candidates were between the ages of 16 and 25 (66%), 9% were under 16 and 7% were over 30.

Within the FCE, an overall mark was assigned to each document on a scale from zero to 40. The rating criteria were mainly based on the learners' success in the communicative task they had to complete, with special emphasis on text organization and cohesion, clear layout, appropriate register and control, accuracy and range of language. Scores were scattered across the scale with a mean of 27.92 and a standard deviation of 5.34. Yannakoudakis et al. (2011a) estimated the reliability of the assessments of 97 documents by comparing them to the ratings of four experts, and obtained an average Pearson correlation of .80 ($p < .0001$).

In addition to containing data about a large number of learners, this dataset had the advantage of being freely available for non-commercial research and educational purposes (Yannakoudakis et al., 2011b), including both the texts and the ratings. It is therefore particularly relevant for the replication objective of this study since other researchers can use it to compare the effectiveness of their formulaic measures to the collgram ones. However, one of its characteristics cannot be overlooked, namely that each learner's document is composed of two texts which may be of different genres. This factor can certainly influence the lexical diversity measures studied here. A document composed of two texts of different genres is likely to have a higher lexical diversity than a document composed of two texts of the same genre, although in both cases the topics of the two texts are different. For this reason, each of the lexical richness and formulaic measures reported

[1] Nine documents were disregarded from the original dataset because they only contained a single text.

below was computed separately for each text of a learner document and then averaged to assign a global score to the document.

4.1.2. The ICLE dataset

This dataset, described in detail by Thewissen (2013) and Granger and Bestgen (2014), was composed of 223 essays extracted from the International Corpus of Learner English (ICLE; Granger, Dagneaux, & Meunier, 2002), representing a corpus of essays written by intermediate to advanced learners of English as a foreign language. To be selected, an essay had to be 500 to 900 words long, argumentative in nature and written by native speakers of French (N=74), German (N=71) or Spanish (N=78). These L1 backgrounds were selected because they were the languages with which the researcher, who linguistically annotated these texts, was the most familiar, or about which she could easily consult native speakers (Thewissen, 2013). The dataset amounted to 151,448 words. Thirty-eight different prompts were used to collect the texts, with half of them only being used once or twice, although some were used much more often, such as *Some people say that in our modern world, dominated by science and technology and industrialisation, there is no longer a place for dreaming and imagination. What is your opinion?* (13%) or *Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television* (11%). Eighty-five percent of the texts were not produced under any temporal constraints and were not intended to be used for evaluation. The learners were all undergraduate students, mainly between the ages of 21 and 25 (66%), and none of them were under 19 or over 31 years of age.

Unlike the texts included in the first dataset, the ICLE essays were not written as part of a standardized procedure for assessing L2 competence. The 223 selected essays were evaluated by two professional raters, who assigned a score based on the CEFR descriptors (Council of Europe, 2001) provided to them. The raters were asked to judge the essay as being one of B1, B2, C1 or C2 (i.e., intermediate [B] and advanced [C] levels of proficiency in the CEFR), carrying out the rating procedure completely independently. They were allowed to use + or – signs to further distinguish between sublevels. These categories were recoded as a numerical scale from 1 (for B1–) to 11 (for C2). The Pearson correlation coefficient between the grades given by the two raters on this 11-point scale was .69 ($p < .0001$, $N = 223$), corresponding to a reliability, estimated by the Spearman–Brown prediction formula, of .82, which is a substantial level of interrater reliability for this kind of task (Artstein & Poesio, 2008). The 34 texts (15%) on which the two raters disagreed by more than one band score (e.g., B1-C1) were submitted to a third rater. The final score of a text was an average of the two (or three) available scores. The dataset contains 66 B1 texts, 62 B2 texts, 67 C1 texts and 28 C2 texts.

4.1.3. Pre-processing

Spelling errors in the learner texts were corrected on the basis of the manual annotations available in both datasets. This is a recommended step before computing the lexical richness measures (Kojima & Yamashita, 2014) and it allows to group instances of bigrams that differ only in one or two minor spelling errors (e.g., *private correspondance* and *private correspondence*). Words were only corrected when there was no doubt about the targeted form. The texts were then tagged for part-of-speech using CLAWS7 (http://ucrel.lancs.ac.uk/claws/). After this step, proper nouns, numbers and non-English words were identified using the part-of-speech tags and the list of unknown terms provided by WMatrix (http://ucrel.lancs.ac.uk/wmatrix/), then they were deleted.

4.2. Formulaic measures

Formulaic measures were calculated as described in Bestgen and Granger (2014). Firstly, all bigrams were extracted from each learner text. They were then looked up in the British National Corpus (BNC), a 100-million-word collection of samples of written and spoken

language designed to represent a wide cross-section of British English from the latter part of the 20th century. The BNC was chosen because British English is the dominant variety of English taught to the majority of the learners in the two analysed datasets. Every bigram extracted from the learner texts that was found in the BNC was assigned two association scores (MI and t-score), which were calculated using the formulas reported in Evert (2009; see also section 2.1). These association scores were used to calculate the first two formulaic measures, i.e., the mean MI score and mean t-score of the bigrams in a text. The third measure, Pabsent, corresponded to the proportion of bigrams from the learner text that were absent from the reference corpus. These three formulaic measures were calculated on the basis of the different bigrams present in each text (types), in order to give more weight to their diversity (Durrant & Schmitt, 2009).

4.3. Lexical richness measures

4.3.1. Lexical diversity

Among the many measures of lexical diversity, the three that McCarthy and Jarvis (2010) recommended in their comparative study were selected because they provide complementary views on lexical diversity and are less influenced by text length than the type-token ratio (even if, according to Treffers-Daller (2013), they are not completely immune to this factor): 1) the Maas index, a log-corrected type-token ratio, which is reversed (1-Maas) to give higher scores to more diversified texts; 2) the Measure of Textual Lexical Diversity (MTLD), which is equal to the mean length of sequential word strings in a text maintaining a type-token ratio of 0.72, and 3) HD-D, which is McCarthy and Jarvis' equivalent of the well-known D measure (Malvern & Richards, 2002), based on the probability of new words being introduced into increasingly long samples of text.

4.3.2. Lexical sophistication

Three lexical sophistication measures were also selected. The first two were derived from the Lexical Frequency Profile developed by Laufer and Nation (1995). These measures corresponded to the proportion of different words considered as advanced, compared to the total number of different words in the text. For the first index, i.e., Beyond 2000 (B2000), words were considered as advanced if they were not in the 2,000 most frequent English words in the General Service List (West, 1953); the second index (not-in-any-list – NIAL) concerned advanced words that also could not be found in the Academic Word Lists (Coxhead, 2000). This second measure was considered particularly relevant for analysing texts written by advanced learners (Nation, 2001). The third measure, S, was recently proposed by Kojima and Yamashita (2014) to enable the lexical richness of texts of limited size to be reliably estimated. S corresponds to the word frequency value, where text coverage is expected to reach 100%. To calculate this, Kojima and Yamashita's (2014) lists, which are based on the spoken section of the BNC, were used.

4.4. Statistical analyses

Relations between pairs of variables were estimated by means of the Pearson correlation coefficient (r). In each of these analyses, the alpha value for deciding whether a correlation is statistically significant was adjusted using the Bonferroni correction. Cohen's guidelines for the social sciences (Cohen, 1988) were followed in order to gauge the importance of the correlations: namely, a medium effect size when r is close to .30, and large when it reaches .50. The t test proposed by Steiger (1980) was used for testing the difference between two non-independent correlations (i.e., two correlations with one common variable), using a significance threshold of .05.

Multiple linear regressions were applied when several variables were used together to predict text quality. They were hierarchical regression models, which allowed for comparing models defined *a priori* in terms of how relevant they were to answering the research questions on the basis of the percentage of variance (R^2) of the text quality scores

that they could explain (see Treffers-Daller et al. (2016) for the use of this statistical technique for comparing several lexical diversity measures). The analyses were first conducted using all of the data available in order to obtain the most complete picture possible. However, to reduce the risk of a positively biased estimate of the effectiveness of the predictors, the models were also assessed using a cross-validation procedure. This was done by randomly dividing each dataset into three as equally sized parts as possible, and by repeatedly using two of these parts to build the model for predicting the scores of the observations in the third part. The randomization procedure was repeated five times and the seed of the random generator was varied in each instance. The R^2 value reported for this procedure was the average of the 15 obtained R^2 values. It is important to note that, although performing cross-validation on a dataset allows researchers to judge the possibility of generalizing the results to another sample extracted from the same population, it does not provide any guarantee that these results can be generalized to another population. Obtaining similar results in several datasets is the only way to substantiate such a claim.

An automatic method of selecting the predictors such as stepwise regression was not used. This is because it has the drawback that, when two predictors explain a common part of the dependent variable, only the best of them (even if it is by a very small amount), is included in the model, while the usefulness of the other is completely masked (Howell, 2013). In other words, stepwise regression does not allow an estimation of the specific proportion of variance accounted for by a particular measure.

Finally, it is worth noting that the partial multicollinearity between predictors, which was expected since several measures of the same construct were simultaneously analysed, posed no problem here; it only affected the stability of the partial regression coefficients rather than the overall fit of the model, which was the criterion used in this study. Furthermore, an analysis of the *variance inflation factors* (VIF) of the models, including all of the predictors, indicated no significant problems according to standard criteria (Hair, Anderson, Tatham, & Black, 1995). In other words, no VIF was close to 10 and only one was greater than five (i.e., 5.5 for S in the ICLE dataset).

5. Results and discussion

5.1. Correlations between formulaic and lexical richness measures and text quality score

Table 1 reports the correlations between the text quality score and each of the formulaic and lexical richness measures. For both datasets, the most correlated measure was the mean MI score. It was statistically significantly more correlated than all other measures, except when compared to the Maas index in the ICLE dataset ($p > .10$). The other two formulaic measures, mean t and Pabsent, were only significantly correlated with text quality in the FCE dataset. For the lexical diversity measures, correlations with the Maas index were larger than those of the other two measures, although the differences were only statistically significant in the FCE dataset. The three lexical sophistication measures were also significantly correlated with text quality in both corpora. While there was no statistically significant difference between the measures in the FCE dataset, the NIAL measure was significantly better than the other two in the ICLE dataset.

Table 1: *Correlations between formulaic and lexical richness measures and text quality*

	Formulaicity			Diversity			Sophistication		
	Mean MI	Mean t	Pabsent	Maas	MTLD	HD-D	B2000	NIAL	S
FCE	.46**	.10**	-.21**	.23**	.19**	.16**	.13**	.11**	.12**
ICLE	.60**	.03	.11	.51**	.44**	.46**	.28**	.48**	.35**

Note. * $p < .05$, ** $p < .01$. $N = 1235$ for the FCE dataset and $N = 223$ for the ICLE dataset.

Regarding the differences between the datasets, seven out of nine measures were much more correlated (in absolute value) with text quality in the ICLE than in the FCE dataset, with the two exceptions being the mean t-score and Pabsent. The potential origins of these differences are discussed in Section 5.4.

These results confirm the existence of an important relationship between the mean MI score and text quality, but the relationship is much weaker for the other two formulaic measures. The results also indicate that measures of lexical diversity and sophistication were related to text quality, although less so than the best formulaic measure.

Table 2: *Correlations between formulaic and lexical richness measures for both datasets*

FCE	Mean MI	Mean t	Pabsent	Maas	MTLD	HD-D	B2000	NIAL
Mean MI	—							
Mean t	.43**	—						
Pabsent	-.23**	-.18**	—					
Maas	.31**	-.09**	.41**	—				
MTLD	.25**	-.09**	.15**	.83**	—			
HD-D	.21**	-.16**	.17**	.86**	.87**	—		
B2000	.00	-.11**	.43**	.29**	.23**	.25**	—	
NIAL	.02	-.10**	.41**	.20**	.13**	.16**	.83**	—
S	.08**	-.13**	.51**	.34**	.27**	.28**	.74**	.70**

ICLE	Mean MI	Mean t	Pabsent	Maas	MTLD	HD-D	B2000	NIAL
Mean MI	—							
Mean t	.37**	—						
Pabsent	-.01	-.39**	—					
Maas	.41**	-.26**	.54**	—				
MTLD	.35**	-.21	.55**	.81**	—			
HD-D	.34**	-.11	.36**	.71**	.80**	—		
B2000	.27**	-.29**	.62**	.60**	.41**	.20	—	
NIAL	.40**	-.21	.69**	.65**	.54**	.39**	.75**	—
S	.36**	-.30**	.74**	.63**	.54**	.31**	.80**	.81**

Note. * $p < .05$, ** $p < .01$. $N = 1235$ for the FCE dataset and $N = 223$ for the ICLE dataset.

5.2. Correlations between formulaic and lexical richness measures

In order to evaluate the relationship between all these measures, the formulaic and single-word measures were correlated two by two in both datasets. These correlations are reported in Table 2. As expected, the strongest correlations were observed between the three measures of lexical diversity and between the three measures of lexical sophistication. Such high correlations were not observed for the three formulaic measures, which was an anticipated result because these measures are intended to provide different views on the learner's formulaic competence.

Table 2 also shows that Pabsent was relatively well correlated with the lexical sophistication measures, especially in the ICLE dataset. Since a high score on a sophistication measure arises from the use of advanced (or rare) words, it suggests that Pabsent highlights the use of bigrams including advanced words that have less chance of occurring in the reference corpus. A contrast between the mean MI score and the mean t-score is also indicated, insofar as the former was mostly positively correlated with lexical richness measures, while the correlations were negative for the latter. This result may be related to the tendency of MI to favour expressions consisting of rare words, which are sophisticated and seldom repeated in a text, while the t-score favours frequent sequences, which are usually composed of common and repeated words. Many correlations were higher in the ICLE than in the FCE dataset and the differences were often quite large.

5.3. Comparing the efficiency of formulaic and lexical richness measures in predicting text quality

The analyses reported above show that formulaic and lexical richness measures predict a statistically significant part of the text quality score, while also being correlated. Therefore, it is necessary to check whether these measures made any contribution of their own to the prediction or whether some measures were of little use, as the portion of the variance in the quality score explained by them is also explained by other measures. To this end, a series of hierarchical linear regressions was performed using the quality score of the texts as the dependent variable and the three groups of measures (i.e., formulaicity, lexical diversity and lexical sophistication), which are in turn each made up of three measures, as predictors. In these analyses, the three measures of the same group were simultaneously introduced into the models in order to obtain the best possible prediction each time, the portion of the variance (or R^2) explained together by the measures, but also the portion specific to each of the three measures being added up.

The R^2 values reported in the columns entitled "Whole dataset" in Table 3 were obtained by estimating the models for all of the data, while those reported in the two rightmost columns were obtained through a cross-validation procedure. Before going into the details, it is interesting to note that the R^2 values obtained using both approaches were very close. As expected, the latter were lower than the former, although the differences were small and did not alter the conclusions. It can thus safely be concluded that the predictive model built on the whole dataset should also fit new samples extracted from the same population.

The first section of Table 3 shows the portions of the variance explained by the different sets of predictors, either alone or in combination. When used alone, the three sets were useful in predicting the quality of the texts. The formulaic measures were much more effective than the lexical diversity measures, which in turn were more effective than the lexical sophistication measures. Combining the two sets of lexical richness measures allowed a better prediction than when used in isolation, especially in the ICLE dataset. However, it did not allow them to achieve the level of efficacy associated with the formulaic measures. The models including the three sets of predictors explained 28 to 52% of the variance of the text quality scores, representing a percentage of 4 to 13% higher compared to that explained by the formulaic measures alone, confirming the interest in combining single-word and multi-word measures.

Table 3

R^2 *values from the hierarchical regression analyses using the three sets of measures as predictors of text quality*

				Whole dataset		CV	
	Formulaicity	Diversity	Sophistication	FCE	ICLE	FCE	ICLE
Overall contribution (Full model R^2)							
	X			.242	.396	.239	.390
		X		.062	.279	.061	.274
			X	.018	.247	.014	.236
		X	X	.067	.347	.059	.316
	X	X	X	.283	.522	.273	.475
Individual contribution (Semi-partial R^2)							
	X			.216	.174	.214	.160
		X		.005	.053	.002	.031
			X	.030	.058	.026	.045
		X	X	.041	.126	.034	.085

Note. An "X" in a column means that the group of measures indicated by the column label was used as predictors. $N = 1235$ for the FCE dataset and $N = 223$ for the ICLE dataset. All R^2 values are statistically significant at the .0001 level, except for the individual contribution of lexical diversity in the FCE dataset, which is significant at the .05 level.

The second question that Table 3 helps to answer focuses on the individual contribution of each set of measures in predicting text quality scores. This individual contribution is calculated using the difference between the R^2 value of the full model and that of the model in which the target set of predictors is not included. For instance, for the FCE whole dataset analysis, R^2 value for the full model was .283, while the model excluding formulaic measures obtained an R^2 value of .067. The difference, .216, is the portion of the variance explained by the formulaic measures alone. This value is provided in the second section of Table 3, together with those for both sets of lexical richness measures. As can be observed, the formulaic measures[2] made a much larger specific contribution than the lexical richness measures, whether considered in isolation or together, especially in the FCE dataset.

All of the measures clearly performed better in the ICLE than in the FCE dataset. This was especially the case for the lexical richness measures, which explained a small portion of the variance in the FCE dataset but explained a quarter each and a third when combined in the ICLE dataset. Note that this difference between the datasets was also observed in the cross-validation results. If the study had been conducted using cross-

[2] Given that the mean MI score is much more correlated with text quality than the other two formulaic measures (see Table 1), and given the large correlations between Pabsent and the lexical sophistication measures (see Table 2), the same regressions were performed by using the mean MI score alone in the formulaic "set". This reduces the R^2 value of all of the models, which include the formulaic measure as a predictor, by approximately 0.04, but does not increase any of the contributions made by the lexical richness measures.

validation on only one of the two datasets, it would have been concluded that lexical richness measures were either good predictors (ICLE) or poor predictors (FCE). Cross-validation does not provide the same information as the analysis of different datasets. This observation confirms the increasingly emphasized need in applied linguistics to replicate analyses on several datasets (Porte, 2012).

Table 4
Comparison of the means for formulaic and lexical richness measures between the two datasets

	FCE		ICLE				
	M	*SD*	*M*	*SD*	*t*	*df*	Hedges' *g*
Mean MI	2.30	0.21	2.21	0.20	4.07	320	0.40
Mean t	48.54	6.74	35.88	5.26	22.64	348	1.94
Pabsent	0.04	0.02	0.06	0.03	12.84	257	1.31
Maas	0.979	0.003	0.981	0.003	7.74	288	1.34
MTLD	68.61	14.06	86.18	21.53	10.79	261	1.14
HD-D	0.80	0.02	0.84	0.02	10.14	319	1.68
B2000	0.07	0.03	0.18	0.05	28.21	247	3.60
NIAL	0.05	0.02	0.10	0.04	16.72	238	2.45
S	2570.02	378.83	3847.07	1191.40	16.14	229	2.20

Note. *M* = Mean. *SD* = Standard Deviation. *t* = Welch's *t* test. *df* = degrees of freedom (adjusted by Welch's test and rounded down to the nearest integer). All *t*s were statistically significant at the .001 threshold.

5.4. Comparison of results according to the datasets

The analyses reported above have highlighted a series of differences between the two datasets. In particular, correlations and R^2 values are almost always higher in the ICLE than in the FCE dataset. Significant differences also exist in the mean scores for each of the measures analysed, as shown in Table 4. These values were compared using Welch's *t* tests for unequal variance and all were found to be statistically significant at the .001 threshold, largely because of the large sample size being compared, a factor known for boosting statistical significance. More meaningful is the fact that almost all effect sizes, measured by Hedges' *g*, were large ($g \geq 0.80$). The ICLE texts are, on average, lexically richer than the FCE texts. This is observed for all measures of lexical diversity and sophistication. By contrast, the FCE texts have, on average, a higher mean MI score, which is the best predictor of text quality and is positively correlated with it. This dissociation between the two types of measures in the two datasets was unexpected. It should, however, be noted that the effect size for the mean MI score is low (unlike that for the mean t-score, but this measure is much less correlated with text quality).

Determining which factor or set of factors was responsible for disparities in the results between the two datasets is not an easy task. As indicated in the Material section, these datasets differ in many factors, such as the age and L1 of the participants, text length, genre, prompt, production situation and even the evaluation criteria. While it is difficult to argue that some of these variables could not have affected the results, three of them seem to merit special attention on the basis of the literature.

First, the text samples in the two datasets are very different since every ICLE learner wrote an argumentative text while two texts of various genres were available for the FCE learners. Even though the texts in the ICLE dataset were obtained on the basis of a

large number of different prompts, they were undoubtedly more homogeneous than those in the FCE dataset. This factor may have played a major role in obtaining higher correlations in the ICLE dataset. The fact that the FCE dataset is freely available could allow other researchers to compare the results obtained with it with those obtained on the basis of datasets less homogeneous than the ICLE and thus to study this factor in detail. It should also be remembered that compared to the FCE dataset, more data were available in the ICLE dataset to assess each of the linguistic variables since the texts in the ICLE were on average 669 words long (*SD* = 93) while the two texts available for each FCE learner amounted only to 373 words (*SD* = 35). Van Hout and Vermeer (2007) showed that the reliability of lexical diversity measures was strongly affected by the number of words available for estimation, with the shortest samples producing the least precise estimates. With respect to the formulaic measures, Durrant and Schmitt (2009) observed larger differences between native and non-native speakers when analysing long texts as opposed to short ones. However, the longest texts they analysed contained more than 3,000 words and the shortest ones less than 600, a significantly greater range than in the analyses reported here.

Another difference that could have affected the results is the fact that the FCE dataset was a result of a fee-paying, recognized examination, unlike the ICLE dataset. In a formal examination, participants tend to take as few risks as possible, favouring accuracy over creativity (Neumann, 2014). This aversion to risk could have led the FCE learners to use simpler and more frequent vocabulary and to avoid combinations of words of which they were not completely sure. This could explain the single-word lexical differences between the two datasets, but not the highest MI score for the FCE dataset. Moreover, the impact of this factor on correlations is not evident. It is still possible that it more strongly affected the best learners, who are more likely to use rare words and complex word sequences, than the weaker ones, thus reducing the differences between them in the FCE dataset.

A third difference that must be emphasized is that the L1s of the learners are much more diverse in the FCE than in the ICLE dataset. Leńko-Szymańska (2014) observed a gradual growth in the frequency of lexical bundles from grades six to 12 for Spanish, Polish and Austrian learners of English, but not for other L1 groups (Japanese, Israeli and Taiwanese). In addition, several studies have highlighted the impact of L1 transfer on the use of collocations and lexical bundles (Nesselhauf, 2005; Parkinson, 2015). In terms of lexical diversity, Yu (2010) did not find any significant correlation with the holistic quality of the texts for the two largest L1 groups (Chinese and Filipino) he studied, but found a correlation for the rest of his sample, consisting of 36 different L1 backgrounds. Moreover, Treffers-Daller et al. (2016) pointed out that learners whose L1 is similar to the L2 might obtain higher lexical diversity scores than learners whose L1 differs more greatly.

It is important to emphasize that, despite the differences discussed above, the research questions were answered in the same way in the two datasets. It is nevertheless regrettable that the origin of these differences is not fully understood. Technically, it would be possible to carry out analyses in the manner of Yu (2010) by contrasting sub-samples of a dataset selected on the basis of a variable, such as text length or L1. This procedure would greatly reduce the amount of data available for estimating the predictive models. More reliable information would be obtained through a study aiming precisely at evaluating the impact of a given factor and whose material would have been selected for that purpose.

6. Conclusion and implications

This study's main objective was to compare the usefulness of formulaic and lexical richness measures when assessing the quality of an L2 learner text. Analyses showed that formulaic measures were the best predictors and that they provided a much higher specific

contribution to the prediction than single-word lexical measures. Of these measures, the mean MI score was clearly found to be the most effective. These results thus answer positively the first research question, by emphasizing the existence of a strong link between formulaic competence, at least as it is measured here, and raters' judgements on text quality, as well as the second research question, by showing the additional benefits formulaic measures bring when compared to single-word lexical measures.

Regarding this second research question, the analyses confirmed the findings of numerous studies concerning a link between the lexical richness of a text and its rated quality (Engber, 1995; Jarvis & Daller, 2013; Treffers-Daller et al., 2016; Yu, 2010). Multiple regression analyses indicated that combining diversity and sophistication measures allows to explain a statistically significant portion of the variance in quality scores, even when the formulaic measures are taken into account. This study, then, argues in favour of using both single-word and multi-word measures for tracking the development of L2 competence in texts.

The findings summarized above were obtained from both datasets, showing that they are replicable. However, several differences in results between the two datasets were observed such as the fact that almost all of the measures were more effective in the ICLE dataset. Further research on better controlled datasets is needed to determine which factors explain the observed differences. Such a study could also help to understand some unexpected results in the literature such as the absence of a relationship between L2 language proficiency and both formulaic competence and lexical richness (Vedder & Benigno, 2016). However, there are many more differences between the various studies into the formulaic competence of L2 learners than between the datasets analysed here, such as oral or written modality, learnt language and, in particular, the types of expressions analysed. The formulaic measures evaluated in this study are calculated automatically and thus "objective "except for the choice of the reference corpus. Although they can be easily applied to many texts, the information they provide is less rich than the qualitative analyses of FSs at the heart of many L2 studies (Henrik Sen, 2013). Combining the two types of approaches, at least in a small study, would allow for a better appreciation of the advantages and disadvantages of each one and lead to some useful modifications to the automatic approach.

A potentially important limitation of the formulaic measures evaluated is that they are only based on bigrams, while many FSs are longer. Extending the approach to these sequences, however, requires a thorough study of the association measures for longer n-grams, as they have received much less attention (Evert, 2009) and as the usual extension of the MI (but also of the t-score) has been criticized (Appel & Trofimovich, 2017). This will also require the automation of procedures designed to combine overlapping n-grams (Appel & Wood, 2016; Chen & Baker, 2016). Another interesting development would be to apply the formulaic measures to longitudinal data. This would involve a relatively large number of texts per learner (Li & Schmitt, 2009) and a relatively large number of learners (Siyanova-Chanturia, 2015). Such a study would facilitate the determination of whether changes in formulaic competence are relatively continuous, or whether this skill is mostly improving after a sufficiently advanced level has been reached or after sufficient exposure to L2 formulaic language, which is considered as a major determining force in FS acquisition (Vedder & Benigno, 2016).

The educational implications of this research lie first and foremost in the field of automatic learner text assessment. Evaluating the quality of texts is essential in foreign language learning, where being able to write good texts is one of the main objectives (Weigle, 2013). In this field, many automated assessment systems such as e-Raters (Educational Testing Service – ETS) have been developed that rely on lexical, syntactic and structural features correlated with writing quality (Burstein, Chodorow, & Leacock, 2004). These systems are used in classrooms to give learners, under the supervision of the

instructor, an initial assessment of the quality of their writings but also to point out particularly well-written passages and potentially problematic points (Ramineni & Williamson, 2013). In these systems, FSs have been largely neglected (Higgins, Ramineni, & Zechner, 2015). It follows that FSs cannot be taken into account during the automatic assessment and that no feedback can be provided about them by the systems. Several approaches which take FSs into account are possible, such as use of n-grams, lexical bundles and collgrams. In future work, it will therefore be interesting to compare their effectiveness and to work out how to combine their useful specificities.

This study supports the current trend in foreign language teaching that highlights the role played by formulaicity in the development of L2 writing and its usefulness for tracking this development. It could be thought that second language teachers ought to be already convinced. Yet in her review, Henriksen (2013) stresses that "Many teachers tend to focus on individual words (e.g., in glosses and tasks) and often lack useful materials for raising learners' awareness of collocations" (p. 41). It is hoped that, in association with the observations of Crossley et al. (2015) and Kyle and Crossley (2015), the results presented above will help to strengthen this trend. More specifically, this study suggests that it may be important for English learners to focus on how (relatively) rare words are combined (those that have a high MI score). The recent implementation of the collgram procedure[3] by Leńko-Szymańska and Wolk (2016) is an important development. It allows users not only to score texts for formulaicity but also to automatically find in a L2 text the most and least native-like bigrams (according to the MI and the t-score) that could then be used for classroom exercises. The collgram calculator could also highlight in reading materials the lower-frequency, but strongly-associated, word sequences characterized by high MI scores. These are particularly difficult to master for L2 learners (Durrant & Schmitt, 2009). However, further studies are needed to qualitatively analyse the outputs of the collgram procedure, and not quantitatively and globally as it is the case here. The present study only raises a tiny part of the veil that covers the potentialities of automatic approaches to identifying FSs.

Finally, the many differences observed between the two datasets, although they do not call into question the general conclusions of the study, should draw the attention of researchers and practitioners to the difficulty inherent in any attempt at generalizing the results of a study to other learner populations and learning contexts (Coxhead, 2015). As stressed by the Language Teaching Review Panel (2008), "The potential for replicating studies in order to validate results is a requirement of scientific enquiry and should become more prominent in establishing and confirming the outcomes of research in L2 learning and teaching" (p. 1). This issue is clearly not unique to formulaic studies, but is probably more acute in this field because, in addition to the usual differences in the learner population, the task and the learning context, there are often methodological differences between the procedures used to identify FSs. Automatic procedures, which are or may soon become available, are a possible response to these difficulties. Developing, improving and evaluating them therefore seem promising.

References

Ackermann, K., & Chen, Y. (2013). Developing the academic collocation list (ACL): A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12*, 235–247.

[3] It is noteworthy that the publicly available implementation of the collgram procedure is based on a different reference corpus (COCA) than the one used here (BNC). This difference does not seem important since Bestgen and Granger (2015) observed that using any of these two corpora in the collgram approach leads to very similar results.

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*, 81–92.
Appel, R., & Trofimovich, P. (2017). Transitional probability predicts native and non-native use of formulaic sequences. *International Journal of Applied Linguistics*, *27*, 24–43.
Appel, R., & Wood, D. (2016). Recurrent word combinations in EAP test-taker writing: Differences between high- and low-proficiency levels. *Language Assessment Quarterly, 13*, 55–71.
Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*, 555–596.
Bestgen, Y. (in press). Evaluating the frequency threshold for selecting lexical bundles by means of an extension of the Fisher's exact test. *Corpora.*
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.
Bestgen, Yves, & Granger, Sylviane (2015, September). *Using collgrams to assess L2 phraseological development: A replication study*. Paper presented at the 3rd Learner Corpus Research Conference, Nijmegen, NL.
Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. London: Longman.
Boers, F., & Lindstromberg, S. (2012). Experimental and intervention studies on formulaic sequences in a second language. *Annual Review of Applied Linguistics*, *32*, 83–110.
Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research*, *10*, 245–261.
Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The criterion online writing service. A*I Magazine*, *25*, 27–36.
Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL, 5*, 31–64.
Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, *14*, 30–49.
Chen, Y., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistic*, *37*, 849–880.
Church, K., Gale, W.A., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisition: Using on-line resources to build a lexicon* (pp. 115–164). Mahwah: Lawrence Erlbaum.
Clear, J. (1993) From Firth principles: Computational tools for the study of collocation. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 271–292). Amsterdam: John Benjamins.
Cobb, T. (2013). Frequency 2.0: Incorporating homoforms and multiword units in pedagogical frequency lists. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 79–108). Amsterdam: Eurosla.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah: Lawrence Erlbaum.
Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics*, *2*, 223–235.
Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238.

Coxhead, A. (2015). Replication research in pedagogical approaches to formulaic sequences: Jones & Haywood (2004) and Alali & Schmitt (2012). *Language Teaching*. Advance online publication. doi:10.1017/S0261444815000221.

Crossley, S.A., Cai, Z., & McNamara, D.S. (2012). Syntagmatic, paradigmatic, and automatic n-gram approaches to assessing essay quality. In P. McCarthy & M. Youngblood (Eds.), *Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference* (pp. 214–219). Menlo Park, CA: The AAAI Press.

Crossley, S.A., Cobb, T., & McNamara, D.S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System*, *41*, 965–981.

Crossley, S.A., Salsbury, T., & McNamara, D.S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*, 240–260.

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.

Daller M., & Xue, H. (2009). Vocabulary knowledge and academic success: A study of Chinese students in UK higher education. In B. Richards, M. Daller, D. Malvern, P. Meara, J. Milton, & J, Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition: The interface between theory and application* (pp. 179–193). Basingstoke: Palgrave Macmillan.

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

Engber, C.A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*, 139–155.

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations* (Unpublished doctoral dissertation). University of Stuttgart, Stuttgart.

Evert, S. (2009). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook* (pp. 1211–1248). Berlin: Mouton de Gruyter.

Forsberg Lundell, F., & Lindqvist, C. (2012). Vocabulary development in advanced L2 French: Do formulaic sequences and lexical richness develop at the same rate? *LIA: Language, Interaction and Acquisition*, *3*, 73–92.

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

Granger S., Dagneaux E. & Meunier F. (2002). *International corpus of learner English*. Handbook and CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.

Gyllstad, H. (2007). *Testing English collocations: Developing tests for use with advanced Swedish learners* (Unpublished doctoral dissertation). Lund University, Lund.

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis* (3rd ed.). New York: Macmillan.

Henriksen, B. (2013). Research on L2 learners' collocational competence and development: A progress report. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 29–56). Amsterdam: Eurosla.

Higgins, D., Ramineni, C., & Zechner, K., (2015). Learner corpora and automated scoring. In S. Granger, G. Gilquin, & F. Meunier (Eds), *Cambridge handbook of learner corpus research* (pp. 567–586). Cambridge: Cambridge University Press.

Howell, D. (2013). *Statistical methods for psychology* (8th edition). Boston: Cengage Learning.

Huang, K. (2015). More does not mean better: frequency and accuracy analysis of lexical bundles in Chinese EFL learners' essay writing. *System*, *53*, 13–23.
Jarvis, S., & Daller, M. (Eds.). (2013). *Vocabulary knowledge: Human ratings and automated measures*. Amsterdam: John Benjamins.
Kojima, N., & Yamashita, J. (2014). Reliability of lexical richness measures based on word lists in short second language productions, *System*, *42*, 23–33.
Kyle, K., & Crossley, S.A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*, 757–786.
Language Teaching Review Panel (2008). Replication studies in language learning and teaching: Questions and answers. *Language Teaching*, *41*, 1–14.
Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322.
Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*, 647–672.
Leńko-Szymańska, A. (2014). The acquisition of formulaic language by EFL learners: A cross-sectional and cross-linguistic perspective. *International Journal of Corpus Linguistics, 19*, 225–251.
Leńko-Szymańska, A., & Wolk, A. (2016, June). *A corpus-based analysis of the development of phraseological competence in EFL learners using the collgram profile*. Paper presented at the 7th Conference of the Formulaic Language Research Network. Vilnius, LT.
Levitzky-Aviad, T., & Laufer, B. (2013). Lexical properties in the writing of foreign language learners over eight years of study: Single words and collocations. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 109–126). Amsterdam: Eurosla.
Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing*, *18*, 85–102.
Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*, 190–208.
Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, *19*, 85–104.
Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke: Palgrave Macmillan.
McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*, 381–392.
Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.
Neumann, H. (2014). Teacher assessment of grammatical ability in second language academic writing: A case study. *Journal of Second Language Writing, 24*, 83–107.
Oppenheim, N. (2000). The importance of recurrent sequences for nonnative speaker fluency and cognition. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 220–240). Ann Arbor: University of Michigan Press.
Parkinson, J. (2015). Noun-noun collocations in learner writing. *Journal of English for Academic Purposes, 20*, 103–113.
Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–226). London: Longman.
Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources & Evaluation, 44*, 137–158.

Porte, G. (Ed.). (2012). *Replication research in applied linguistics*. Cambridge: Cambridge University Press.
Qi, Y., & Ding, Y. (2011). Use of formulaic sequences in monologues of Chinese EFL learners. *System*, *39*, 164–174.
Qin, J. (2014). Use of formulaic bundles by non-native English graduate writers and published authors in applied linguistics. *System, 42*, 220–231.
Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing, 18*, 25–39.
Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.
Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*, 487–512.
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Šišková, Z. (2012). Lexical richness in EFL students' narratives. *University of Reading Language Studies Working Papers*, *4*, 26–36.
Siyanova-Chanturia, A. (2015). Collocation in beginner learner writing: A longitudinal study. *System*, *53*, 148–160.
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*, 214–225.
Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251.
Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *Modern Language Journal*, *97*, 77–101.
Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French: An exploration of the validity of D, MTLD and HD-D as measures of language ability. In: S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79–104). Amsterdam: John Benjamins.
Treffers-Daller, J., Parslow, P., & Williams, S. (2012, November). *Automated assessment of lexical diversity and n-grams in the Pearson test of English*. Paper presented at the Academic Language Testing Forum, Bristol, UK.
Treffers-Daller, J., Parslow, P., & Williams, S. (2013, March). *Automated assessment of lexical diversity and n-grams in essays at different levels of the CEFR*. Paper presented at the Annual conference of the American Association for Applied Linguistics, Dallas, TX.
Treffers-Daller, J., Parslow, P., & Williams, S. (2016). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*. Advance online publication. doi: 10.1093/applin/amw009
Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers & the Humanities*, *32*, 323–352.
van Hout, R., & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 93–115). Cambridge: Cambridge University Press.
Vedder, I., & Benigno, V. (2016). Lexical richness and collocational competence in second-language writing. *International Review of Applied Linguistics in Language Teaching*, *54*, 23–42.
Verspoor, M., Schmid, M.S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, *21*, 239−263.
Vidakovic, I., & Barker, F. (2010). Use of words and multi-word units in skills for life writing examinations. *Cambridge ESOL: Research Notes*, *41*, 7–14.
Weigle, S.C. (2013). English as a second language writing and automated essay evaluation. In M.D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 36–54). New York: Routledge.

West, M. (1953). *A general service list of English words*. London: Longman.
Wood, D. (2015). *Fundamentals of formulaic language: An introduction*. London/New York: Bloomsbury.
Wood, D., & Appel, R. (2014). Multiword constructions in first year business and engineering university textbooks and EAP textbooks. *Journal of English for Academic Purposes, 15*, 1–13.
Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics*, *32*, 231–254.
Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011a). A new dataset and method for automatically grading ESOL texts. In Y. Matsumoto & R. Mihalcea, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technology* (pp. 180–189). Portland: Association for Computational Linguistics.
[dataset] Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011b). CLC FCE Dataset. https://www.ilexir.co.uk/datasets/index.html.
Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*, 236–259.

Appendix 1: Typical prompts used for collecting the FCE dataset.
A typical prompt for the first task was: "*Your English class is going to spend three days in London. The Principal of your college, Mr Robertson, has already organised the programme. However, the students in your class have seen an advertisement for the London Fashion and Leisure Show and you would all like to go to the show. Your class has asked you to write to Mr Robertson about this. Read the extract from Mr Robertson's programme, the advertisement, and your notes. Then, using the information, write a letter to Mr Robertson.*"

A frequent prompt for a story (15% of all the texts for task 2) found in the dataset is: "*Your teacher has asked you to write a story for the school's English language magazine. The story must begin with the following words: Unfortunately, Pat wasn't very good at keeping secrets. Write your story.* "

A frequent prompt for a composition (also 15% of all the texts for task 2) in the dataset is: *Your class has had a discussion about how science and technology affect our lives. Your teacher has now asked you to write a composition, answering the following question: How has modern technology changed your daily life? Write your composition.*