Accepted for publication in *Language Learning* (Wiley Blackwell)

To what extent are multiword sequences associated with oral fluency?

Parvaneh Tavakoli, University of Reading
Takumi Uchihara, University of Western Ontario

Abstract

This study examines the relationship between oral fluency and use of multiword sequences (MWSs) across four proficiency levels (Low B1 to C1 of the Common European Framework of Reference; Council of Europe, 2001). Data from 56 learners taking the TEEP speaking test were analyzed for different measures of fluency (speed, breakdown, repair) and MWSs (frequency, proportion, association). Results showed that (a) high frequency n-grams correlated positively with articulation rate, (b) n-gram proportion correlated negatively with frequency of mid-clause pauses, and (c) n-gram association strength correlated positively with frequency of end-clause pauses and negatively with repair frequency. The qualitative analysis suggested that the test-takers borrowed some task-specific n-grams from the task instructions and used them frequently in their performance. While lower proficiency speakers used these n-grams verbatim, C1 level speakers used them competently in a variety of forms. Significant implications of the findings for phraseology and language testing research are discussed.

*Key words***:** fluency, multiword sequences, proficiency level, n-grams

Acknowledgement

This project was partly funded by International Study and Language Institute at the University of Reading. We are extremely grateful to the Language Testing Team in general and to Professor David Carter, Professor John Slaght and Gill Kendon for making the data available to us and for their continuous help and support. We would also like to thank Dr Ann-Marie Hunter for her invaluable help with the measurement of fluency in this project and Dr Michael Karas for his constructive feedback on an earlier version of the manuscript. We are grateful to the journal editors and anonymous reviewers for their constructive feedback on earlier drafts of the paper.

Introduction

Oral fluency, the "flow, continuity, automaticity, or smoothness of speech" (Koponen & Riggenbach, 2000, p. 6), is a prime characteristic of successful language communication that has recently become central to many second language (L2) acquisition studies. Investigating fluency helps researchers understand how L2 is processed, produced, and acquired. In addition, exploring fluency enables them to examine some abstract dimensions of SLA, including implicit knowledge, proceduralization of newly learnt structures, and automatization of the speech production processes (Loewen & Sato, 2018; Paradis, 2009; Suzuki & DeKeyser, 2017). For this reason, fluency is often examined in relation to other aspects of language processing and performance. The current study belongs to this body of research as it aims to investigate the relationship between fluency and formulaic use of language. The interest in this relationship is built on the psycholinguistic research evidence that suggests formulaic use of language reduces the amount of language planning and processing, and therefore it facilitates oral production (see Siyanova-Chanturia & Van Lancker Sidtis, 2018 for the review). One line of research in this area (e.g., Nattinger & DeCarrio, 1992; Pawley & Syder, 1983; Wood, 2015; Wray, 2002; see Granger, 2018 for the review) has focused on the use of multiword sequences (MWSs), combinations of words that appear together highly frequently in a target language (Garner & Crossley, 2018), and its relationship with proficiency. The findings of this line of research suggest that more proficient L2 users have a better command of MWSs in terms of range, frequency, and sophistication (e.g., Durrant & Schmitt, 2009; Garner & Crossley, 2018; Kyle & Crossley, 2015; Paquot, 2019). The findings also indicate that use of MWSs is related to oral fluency (Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006; Stengers, Boers, Housen, & Eyckmans, 2011; Tavakoli, 2011; Wood, 2009, 2010). While the existing research evidence highlights the relationship between fluency and MWSs, no study thus far has examined this relationship systematically across a range of assessed levels of proficiency to see if there is a linear relationship between the two. Understanding the relationship between MWSs and fluency at different levels of development is important for SLA as it will help shed light on processes involved in speech production. Similarly, such knowledge can help the language testing discipline to develop a valid assessment of learner phraseological knowledge and its relationship to fluency at different levels of proficiency. This is the gap the current study aims to help fill by investigating, both quantitatively and qualitatively, the relationship between MWSs and fluency from 56 participants at four levels of proficiency assessed via the speaking paper of the Test of English for Educational Purposes (TEEP).

Literature Review

Multiword Sequences

Use of MWSs has recently attracted SLA researchers' attention given its important role in language processing and production. 'MWS' is a catch-all term referring to combinations of words that appear together highly frequently in a target language (Garner & Crossley, 2018; Jeong & Jiang, 2019). MWSs include various forms of lexical strings, such as idioms (e.g., *kick the bucket, spill the beans*), restricted collocations (e.g., *heavy traffic, blow a fuse*), phrasal verbs (e.g., *hung up, call off*), binomials (e.g., *bride and groom, ladies and gentlemen*), proverbs (e.g., *birds of a feather flock together*), and lexical bundles or 'n-grams' as we use in this study (e.g., *and so on, one of the, I don't know*). Since the advancement of corpus-based techniques, n-grams have received increased attention in SLA research. Unlike previous research relying on native-speaker intuition in describing language units, n-gram studies use an objective frequency-based approach to identify recurrent sequences of n (e.g., two, three, four, and longer) consecutive words from learner corpora (Paquot & Granger, 2012). This learner corpora research generally takes two approaches to analyzing MWSs production. In the first approach, n-grams are measured using only text-internal data (i.e., learner corpora) by counting the number of contiguous strings of words of a given length. For example, Huang (2015) extracted recurrent strings of three to five words from argumentative essays written by

two groups of L1 Chinese learners, and compared them in terms of size, range, and accuracy (see Paquot & Granger, 2012, pp. 138-140 for other examples). In the second approach, researchers draw on text-external data (i.e., reference corpora, such as British National Corpus [BNC] and Corpus of Contemporary American English [COCA]) to extract lexical bundles (Garner & Crossley, 2018; Kyle & Crossley, 2015). For word combinations to be qualified as lexical bundles, such combinations found in learner corpora (i.e., L2 speakers) must appear in language used by L1 speakers. In this regard, this approach is concerned with the extent to which L2 learners employ target-like MWSs. In the current study, we adopt the second approach to defining and measuring MWSs with a specific focus on bigrams and trigrams based on earlier studies using COCA as a reference corpus (Garner & Crossley, 2018; Kyle & Crossley, 2015).

Oral Fluency

Fluency is often considered a complex and multifaceted construct that is difficult to define and measure (Lennon, 1990; Segalowitz, 2010). For the purpose of this study, we consider fluency as the general ease, flow, and continuity of speech characterized by temporal and acoustic features, and dysfluency markers. Segalowitz defines fluency in terms of three dimensions of cognitive, utterance, and perceived fluency. Cognitive fluency refers to "the efficiency of operation of the underlying processes responsible for the production of utterances" and is distinguished from utterance fluency, "the features of utterances that reflect the speakers' cognitive fluency", and perceived fluency, "the inferences listeners make about speakers' cognitive fluency based on their perceptions" (Segalowitz, 2010, p. 165). As can be seen, utterance fluency is the only aspect of fluency in this model that can be measured objectively through the analysis of the acoustic characteristics of speech. While the three dimensions of fluency are internally dependent and highly interrelated, in the current study, we are interested in utterance fluency as it allows for an objective measurement of the concept of fluency.

Measuring fluency is complex for a range of theoretical and empirical reasons, including the difficulty to decide which measures best characterize oral proficiency. Skehan (2003) and Tavakoli and Skehan (2005) proposed a triadic framework in which fluency should be measured from three distinct aspects of: a) speed, that is, measures that reflect the flow and continuity of speech, b) breakdown, that is, pauses and silences that break the flow of speech, and c) repair, that is, measures that reflect the monitoring and repair processes such as false starts and reformulations. Researchers have found the triadic framework useful in explaining different aspects of the speech production process. For example, speed fluency is hypothesized to reflect the degree of automaticity (Kahng, 2014; Skehan, 2014), breakdown fluency indicates difficulties at the conceptualization and formulation stages of Levelt's model (explained below, Kormos, 2006), and repair fluency reflects the monitoring processes of the L2 production process (Ahmadian, 2011; Kormos, 1999).

Multiword Sequences and Speech Production

Theoretical accounts explaining the relationship. The important relationship between phraseological knowledge and oral proficiency has been explained by psycholinguistic research evidence. Though it remains debatable whether MWSs are stored and retrieved as holistic units to the same extent as single-word items (Siyanova-Chanturia & Martinez, 2015), psycholinguistic research has suggested that MWSs (e.g., *in the middle of the*) are processed differently from novel strings of language (e.g., *in the front of the*) with processing advantages for the former over the latter in both receptive and productive language tasks (see Siyanova-Chanturia & Van Lancker Sidtis, 2018 for review). The increase in speed of language processing, an advantage that enables speakers to produce utterances more fluently, is linked to freeing up the attentional resources that speakers need to attend to other aspects of language production such as articulation and monitoring (Kormos, 2006; Skehan, 2009).

Levelt's (1989) speech production model, initially proposed for L1 speakers and later refined to explain L2 speakers' speech production process (Kormos, 2006), is helpful in explaining the relationship between use of MWSs and fluency. In this model, speech production involves at least three different stages of *conceptualization*, *formulation* and *articulation*. The conceptualization stage is where a preverbal message is generated. At this phase, the speaker's communicative intention is encoded into a coherent conceptual plan, while the message that is about to be sent is monitored. The preverbal message then moves to the formulation stage where lexical selection and grammatical encoding take place. In this stage, the appropriate lemmas (i.e., lexical items unspecified for phonological form) are activated in the mental lexicon, lemmas are placed into syntactic surface structures, and morphophonological and phonetic encoding are carried out. The product from these stages then moves to the articulation stage where the phonetic plan is executed before speech is produced.

Compared to L1 processing and production, L2 processing and production is known to be challenged by the limitations of the L2 mental lexicon as it is "smaller, less organised, likely slower in access, less elaborated with syntactic and collocational information, and contains a narrower repertoire of formulaic language" (Skehan, Foster & Shum, 2016, p. 98). A larger repertoire of formulaic language is therefore one way to help reduce the load on attentional resources that can be used in parallel processing at different stages of speech production (Kormos 2006; Skehan 2014). More specifically, the benefit of knowledge and use of MWSs for oral fluency can be realized in lexical selection at the formulation stage (Kormos, 2006; Levelt, 1992). In lexical selection, speakers retrieve an appropriate lemma from a myriad of alternatives in the mental lexicon. At this stage, speakers with a large repertoire of MWSs can retrieve longer units of word constituents with similar processing cost as needed for retrieval of single-word items. This would hypothetically allow them to 'buy' processing time in preparation for other processing needs such as syntactic processing and subsequent message generation (Boers et al., 2006; Skehan, 1998). In contrast, speakers with a smaller repertoire of MWSs may not enjoy such a processing advantage because they may exhaust cognitive resources in an attempt to retrieve every constituent item of the whole multiword unit. In addition, it can be hypothesized that phraseologically proficient speakers are less likely to show hesitations or pauses within the sequences because phrase structures are relatively fixed and less likely to be subjected to significant grammatical modification (Boers et al., 2006). Therefore, it can be argued that the fluid functioning of the formulator depends to a great extent on the size and organization of the mental lexicon and mapping of lemmas to concepts. This is where MWSs are purported to help enhance fluency as they facilitate access and retrieval of lexical units and free up attentional resources that are needed to deal with other aspects of speech performance.

Empirical evidence for the relationship. In the field of SLA research, there is a growing interest in investigating the relationship between the use of MWSs and L2 proficiency (Boers & Webb, 2018), which has been motivated by the strong theoretical and empirical psycholinguistic evidence discussed above. However, the majority of studies in this area have focused on written production using an essay writing task (e.g., Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Granger & Bestgen, 2014; Kim, Crossley, & Kyle, 2018; Kyle & Crossley, 2015; Kyle, Crossley, & Berger, 2018; Laufer & Waldman, 2011; Paquot, 2019). The findings of these studies suggest that L2 writers of higher proficiency, compared to lower proficiency L2 writers, tend to have richer phraseological knowledge as well as better control of MWSs in spontaneous communication.

Three recent studies conducted by Crossley and colleagues (Garner & Crossley, 2018; Kim et al., 2018; Kyle & Crossley, 2015) shed light on the relationship between oral proficiency and MWSs. Adopting corpus-based measures to identify recurrent two-word and three-word sequences (bigram and trigram) in terms of proportion, frequency, and associational strength (e.g., T-score, Mutual Information [MI] score), the researchers examined the extent to which L2 speakers employ target-like multiword sequences. A cross-sectional study conducted by Kyle and Crossley (2015) found that holistic oral proficiency scores (human rating) were significantly and positively correlated with a

range of n-gram frequency and proportion scores, indicating that orally proficient speakers produce a greater number of target-like sequences frequently used by L1 speakers. Two longitudinal studies conducted by Kim et al. (2018) and Garner and Crossley (2018) confirmed that L2 speakers produce a gradually increasing proportion of bigrams and trigrams frequently found in an L1 spoken reference corpus. The results also suggested that L2 speakers produce more high frequency bigrams over time. However, L2 speakers showed little or no change in n-gram association scores (T score and MI) with the development of proficiency. This finding is in line with Bestgen and Granger's (2014) study highlighting a lack of longitudinal change in association scores in L2 writing. In sum, the existing literature supports the view that there is a positive relationship between general oral proficiency and use of MWSs.

Some studies have more specifically examined the relationship between oral fluency and formulaic use of language (Boers et al., 2006; Stengers et al., 2011; Tavakoli, 2011; Wood, 2009, 2010). Overall, these studies have shown significant positive relationships between use of MWSs and oral fluency. Boers et al. (2006) elicited speech samples from 17 EFL learners in Belgium through two speaking tasks (i.e., retelling and narrative), which were submitted for fluency assessment using an experienced teacher's rating. The study found a significant and moderate correlation between MWS counts and perceived fluency scores ($r = .393$, $p < .05$). Like Boers et al. (2006), Stengers et al. (2011) elicited speech samples through a retelling task from two groups of L2 learners in Belgium (26 studying English and 34 studying Spanish). The study found medium-to-large correlations between fluency rating scores and MWSs counts for both learners of English ($r = .550$, $p < .01$) and learners of Spanish ($r = .361$, $p < .01$). As opposed to the subjective human rating approach adopted by Boers et al. (2006) and Stengers et al. (2011), Wood (2009, 2010) assessed fluency objectively using multiple temporal measures, including speech rate (i.e., mean number of syllables per minute) and mean length of run (i.e., total number of syllables/words per run in a sample). Working with one Japanese ESL learner in Canada over a period of a six-week fluency instruction course, Wood (2009) reported that fluency improvement took place in parallel with an increase in the number of MWSs. Wood (2010) reported similar findings when examining the development of MWSs of 11 ESL learners over a six-month study in Canada. In line with Wood (2009), the study found a significant improvement in a range of speed fluency measures (e.g., speech rate, mean length of run) which was accompanied by a significant increase in the use of MWSs in speech.

While the emerging research evidence summarized above highlights a positive relationship between oral fluency and MWSs, we find this evidence limited in a number of ways. Firstly, some of these studies were based on relatively small sample sizes ($N = 1$ in Wood, 2009; $N = 11$ in Wood, 2010) or participants with a restricted range of L2 proficiency levels (upper-intermediate to advanced levels in Boers et al., 2006, and CEFR B2 level in Stengers et al., 2011). These limitations in sample size and the restricted range of proficiency might have an impact on the generalisability of the findings. Secondly, the measurement of fluency in earlier studies (Boers et al., 2006; Stengers et al., 2011) was based on subjective judgements of fluency. Given the interest in the field of language testing in moving towards a more objective measurement of L2 ability (Tavakoli, Nakatsuhara & Hunter, 2017), it is important to employ more objective measurements of fluency (Thomson, Boers, & Coxhead, 2017). The few studies that have examined the relationship between objective measures of utterance fluency and MWSs have assessed fluency in a rather limited way by examining only one aspect of fluency, for example, speed fluency (e.g., Wood, 2009).

Research aims and questions

The current study aims to help develop a better understanding of how L2 speakers' use of MWSs relates to their oral fluency at different proficiency levels. The following research questions guide the study:

- To what extent does use of MWSs measured through n-grams (proportion, frequency, and strength of association) relate to level of proficiency in the TEEP speaking test?
- To what extent is use of MWSs measured through n-grams (proportion, frequency, and strength of association) associated with oral fluency (speed, breakdown, and repair) in the TEEP speaking test?

Methodology

TEEP Speaking test

The Test of English for Educational Purposes (TEEP) is a standardized proficiency test used by a number of universities in the UK to obtain information about candidates' proficiency before starting a university degree. The TEEP speaking test includes three tasks, all on the same topic/question. The first task, used to elicit data in the current study, is an extended monologic task in which the test-taker is expected to speak for 3 minutes about a given topic. Unlike some other international tests providing a 30 second planning time (e.g., Aptis), TEEP provides a longer planning opportunity before each speaking test task. This opportunity seems to help reduce the cognitive load and communicative pressure of the task by allowing test-takers to plan for what they want to say. Table 1 below provides some information about the three speaking test tasks within the TEEP test. For reasons of test security, unfortunately, we cannot provide task content or instructions.

Table 1 Structure of the TEEP Speaking test

Part	Task	Mode	Example	Planning time	Response Time
1	Individual talk (role plays)	Monologue	Question: Which is better; private or public services?	4 minutes	3 minutes
2	Scenario discussion	Dialogue	In pair, discuss with your partner and analyze the question	2 minutes	4 minutes
3	Focus question	Further discussion	Discuss the question further with your partner, and agree or disagree!	None	No time limit but generally about 2 mins

The speaking section of the test is rated on a 9-point scale ranging from 0 where the speaker makes no attempt to speak, to level 8 where the speaker is very proficient. Levels 3 and below are called "limited speaker", and the highest level is considered "very good speaker". Drawing on both global and analytic rating scales, the TEEP speaking rating scales assess performance against the following criteria: *explaining ideas*, *interaction*, f*luency*, *accuracy*, *range*, and *intelligibility*. Test-takers' performance is examined by two[i] trained examiners, one acting as an interlocutor and the other as an assessor. The interlocutor introduces the questions and provides guidance before the conversation starts, but she/he does not participate in the conversation. The assessor acts as an observer, sitting quietly at the back of the room listening and examining the test. Interlocutors provide holistic grades, namely explaining ideas, information, and interaction, whereas the examiners give both holistic grades and analytical grades for fluency, accuracy and range, and intelligibility. The examiners and interlocutors work with a set of validated marking scales and marking descriptors for each of the six criteria mentioned above to assess the candidate's performance. The different criteria used for the assessment of speaking receive equal weighting. For further information about the test and to see samples of the past papers, please visit https://www.reading.ac.uk/ISLI/study-in-the-uk/tests/isli-test-teep.aspx. The performances are all recorded by the test administration team.

Data

The data for the current study was provided by the International Study and Language Institute at the University of Reading. The data were 56 samples of test-takers' Task 1 performance which were assessed by two (or at times three) experienced raters and placed on the 9-point rating scales. The test-takers were pre-degree university applicants who took the test as one of the entry requirements to their programmes of study at a British University. The test-takers were all non-native speakers of English from a range of 10 different L1 backgrounds including Chinese, Arabic, Japanese, Kazakh, Thai, Greek, Turkish, and Portuguese. Based on the assessment of their speaking skills, the test-takers' spoken proficiency levels had been rated at four levels of 5.0, 5.5, 6.5, and 7.5, which are equivalent of Low B1, High B1, B2 and C1 at the CEFR level respectively. The proficiency level scores were based on the overall assessment of the test-takers' performance on the TEEP speaking test. The data used for measuring utterance fluency, however, came from their performance of Task 1. This data set comprised 56 task performances, totalling 168 minutes of recordings (i.e. 56 performances x 3 minutes). There were 11 participants at Low B1, 14 at High B1, 17 at B2, and 14 at C1 level of proficiency.

Measuring fluency

Following recent fluency research, we measured utterance fluency in terms of speed, breakdown, and repair fluency (Skehan, 2003; Tavakoli & Skehan, 2005; Kahng, 2014). These three aspects have been shown by different studies as distinct factors underlying the fluency construct (Skehan & Foster, 1997; Kahng, 2014; Tavakoli & Skehan, 2005). For each aspect of fluency, there are several measures that can be used to reflect the speaker's degree of speed (e.g., speech rate, articulation rate, and phonation time ratio), breakdown (e.g., length and frequency of filled and silent pauses) and repair (e.g., frequency of disfluency markers, repetitions, and self-corrections). However, researchers have been warned against using several measures from each aspect as they may correlate and overlap with each other (Bosker et al., 2013; Kahng, 2014; Skehan, 2009). To avoid this problem, we followed the recent research findings in this area to choose our fluency measures.

Articulation rate, that is, total number of syllables divided by total amount of speaking time per minute (excluding pauses), was selected as it is reported to be a reliable representation of speed fluency (de Jong et al., 2015; Kahng, 2014; Mora & Valls-Frerrer, 2012; Suzuki & Kormos, 2019). Following recent research in this area (de Jong et al., 2015; Kahng, 2014; Skehan, 2014; Suzuki & Kormos, 2019), we chose frequency of silent pauses per minute to examine breakdown fluency. We also examined pause location in terms of whether they were in the middle of the clause or at end of the clause. Research in this area (Skehan et al., 2016; Tavakoli, 2011; Tavakoli, Nakatsuhara & Hunter, in press) suggests that L2 speakers, especially at lower proficiency levels, pause more frequently in mid-clause position, whereas native speakers pause more frequently at end-clause junctures. Some emerging evidence also suggests that there is a link between lexical knowledge and pause locations (de Jong, 2016), that is, learners with limited L2 lexical knowledge are more likely to pause both in mid-clause and end-clause positions. Following de Jong and Bosker (2013), a pause in the current study is defined as an instance of silence longer than 250 milliseconds. Finally, following Hunter (2017) and Skehan (2009), the total number of repairs, namely hesitations, repetitions, reformulations, and self-corrections per minute, was used to reflect repair fluency.

Measuring MWSs

There are generally two approaches to defining and measuring MWSs: a phraseological or a frequency-based approach (Boers & Webb, 2018; Granger & Paquot, 2008). A phraseological approach often involves judgements of formulaicity according to a range of identification criteria, including semantic transparency of individual words comprising multiword sequences (e.g., *kick the*

bucket), phonological structure of word strings (e.g., vowel reduction, assimilation), and/or grammatical irregularity (e.g., *if I were you*) (Howarth, 1998; Myles & Cordier, 2017; Wood, 2009, 2010; Wray, 2002). A major issue with this approach is the fact that this method involves a fair degree of subjectivity which might impact on the reliability of the analysis. Notably, earlier studies showed a relatively low interrater agreement of judging units of MWSs between trained judges—r < .60 in Boers et al., 2006 and Stengers et al., 2011—falling far short of the median interrater reliability in SLA research (r = .92; Plonsky & Derrick, 2016). In contrast, a frequency-based approach determines formulaicity of word combinations according to frequency of two or more words co-occurring in an external reference corpus (Gablasova, Brezina, & McEnery, 2017). For instance, the corpus-based approach identifies consecutive and recurrent sequences of a given number of words (i.e., n-grams), including grammatically complete or incomplete word combinations, such as bigrams (e.g., *in the, think that*) and trigrams (e.g., *one of the, the fact that*). With a relatively larger sample size in the current study (N = 56) compared to previous fluency studies (N = 1 to 34 in Boers et al., 2006; Stengers et al., 2011; Wood, 2009, 2010) and a potential reliability issue of human judgements of formulaicity, we adopted a frequency-based measure of recurrent word combinations or n-grams in order to identify MWSs objectively in our data.

N-gram Analysis

Following earlier phraseological research (Garner & Crossley, 2018), we employed three n-gram indices—proportion, frequency, and association—based on two- and three-word contiguous chunks (i.e., bigram and trigram tokens). To calculate n-gram scores, we used the Tool for the Automatic Analysis of Lexical Sophistication 2.0 (TAALES: Kyle & Crossley, 2015; Kyle et al., 2018). TAALES is an up-to-date computational tool used to assess various facets of learner performance in terms of lexical sophistication, such as word frequency, range, and psycholinguistic word information. We selected the spoken subsection of the Corpus of Contemporary American English (COCA: Davies, 2009) as a reference corpus with which to calculate a set of n-gram indices. Our selection of the corpus was based on the relatively large size of its spoken subsection, comprised of 79 million words from spontaneous conversations from a wide range of TV and radio programs in the United States recorded over the last 25 years.

N-gram measures

As mentioned above, the current study used TAALES to calculate three types of n-gram measures (proportion, frequency, and association), producing a total of eight score indices (two proportion, two frequency, and four association). Instead of setting a cut-off point to divide lexical combinations into MWSs and non-MWSs, mean frequency and associate scores were computed. This decision was made with the view that MWSs should be described on a grading scale of frequency and associative strength rather than any dichotomous classification of formulaicity (e.g., frequent vs. infrequent, associated vs. not associated) (Durrant & Schmitt, 2009; Ellis, 2012).

N-gram proportion measures. Proportion score indices refer to the proportion of bigrams and trigrams in learner spoken data that are also found in the representative corpora. In this study, TAALES calculated the proportion of bigrams and trigrams produced in a learner sample based on the 30,000 most frequent n-grams in the spoken subsection of COCA. Higher proportion scores indicate that speakers produce a large number of target-like n-grams while text length is controlled for. Previous research suggests that high proficiency L2 users show a greater proportion of n-grams in their language production (Garner & Crossley, 2018; Kyle & Crossley, 2015).

N-gram frequency measures. Unlike n-gram proportion indices which are based on binary scoring (i.e., presence or absence of learner-produced n-grams in the reference corpora), frequency score indices base their scoring on a continuous scale. They assign a frequency score to all the bigrams and trigrams in a learner text and average all assigned scores to yield a single composite score per speaker. Following Kyle and Crossley's (2015) suggestion, we used logarithmic bigram

and trigram frequency scores instead of raw frequency scores in order to control for Zipfian effects common in word frequency lists. Higher frequency scores indicate that speakers produce a larger number of high frequency target-like n-grams while text length is controlled for. Research in this area suggests that proficient L2 users produce a greater number of high frequency n-grams in their language production (Garner & Crossley, 2018; Kyle & Crossley, 2015) than the less proficient speakers.

N-gram association measures. N-gram association score indices show the strength of association between individual words comprising bigrams or trigrams. Association indices are similar to the proportion and frequency indices in that they are all based on frequencies of word co-occurrence found in a given corpus. However, association measures are less strongly influenced by the individual frequencies of constituent words because they are controlled for in calculating association scores. Of five association measures available in TAALES, we selected two measures, T-score and Mutual Information (MI), that have been extensively used in SLA research investigating MWSs in both spoken and written production (e.g., Durrant & Schmitt, 2009; Garner & Crossley, 2018; Granger & Bestgen, 2014). These two indices are similar in comparing the observed frequency of n-grams in the reference corpus (COCA in this study) with the expected frequency computed on the basis of the frequency of the constituent words (Evert, 2005). However, these two association measures tend to yield insight into different sets of word combinations. MI highlights combinations "which may be less common, but whose component words are not often found apart" (e.g., *ultimate arbiter, tectonic plates*), whereas T-score highlights very frequent combinations, whose rankings "are very similar to rankings based on raw frequency" (e.g., *good example, hard work*) (Durrant & Schmitt, 2009, p. 167). Research has suggested that L2 essays written by high proficiency writers contain word combinations receiving higher MI scores (more sophisticated MWSs) but lower T-scores (fewer easy MWSs) than essays written by low proficiency writers (e.g., Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Granger & Bestgen, 2014). However, no research has so far shown any significant relationship between these association measures and oral proficiency (Garner & Crossley, 2018; Kim et al., 2018).

Data Analysis

The first step in the data analysis was to transcribe the spoken data before coding them for fluency and n-gram measures (see discussion above for the details of the measures). The unpruned transcriptions were segmented into AS-Units (Foster, Tonkyn & Wigglesworth, 2000) and clauses. This enabled us to investigate pauses in mid and end-clause positions. For the analysis of fluency, PRAAT (Boersma & Weenik, 2013) was used to identify the measures of speed and pausing. The data were annotated manually on PRAAT by an expert researcher with extensive experience of analyzing similar data. The annotation process involved several stages including listening to the extracts of speech, inspecting the spectrograms produced in PRAAT and identifying and tagging fluency features (e.g., pauses and repairs) on a corresponding grid.

Although PRAAT measurement is precise and accurate, the entire data set was subjected to PRAAT analysis for a second time to ensure a high level of intra-rater reliability. The repair measures were also coded manually by one of the researchers, and 20% of the transcripts were coded for a second time by another researcher to check the inter-rater reliability. A reliability coefficient of .95 was obtained between the first and second coding for the repair measures. Prior to n-gram analysis using TAALES (Kyle & Crossley, 2015), all the transcripts used for fluency analysis were inspected to fix obvious pronunciation errors and remove orthographic markings of pausing (e.g., *uh, um*), which is a prerequisite for lexical analysis. The resulting transcripts ranged between 93 and 403 words in length (M = 260.8, SD = 72.2).

Results

Before running correlation analyses to answer the research questions, the descriptive statistics of the data set were examined. Table 2 provides the means, medians, standard deviations, maximums, and minimums for the different fluency and n-gram measures across proficiency levels. The figures in this table suggest that the speed of performance increases and length of pause, both mid- and end-clause, decreases as proficiency develops. The picture is less clear regarding total repair, for which a consistent pattern is not observed. As for n-gram units of analysis, the descriptive statistics suggest that frequency, proportion, and association measure of T-score increase as proficiency develops. However, the association measure of MI generally decreases with the development of proficiency. These observations are further discussed in the following sections.

Table 2 Descriptive statistics for all measures of fluency and n-gram (N = 56)

Measures	Levels	Mean (Standard Deviation)	Median	Minimum	Maximum
Articulation rate	5.0	216.75 (24.67)	215.14	187.703	270.06
	5.5	224.99 (17.21)	224.25	199.29	257.34
	6.5	231.77 (17.80)	230.24	203.85	260.59
	7.5	240.11 (25.10)	239.39	193.50	285.29
Frequency of mid-clause pauses	5.0	12.52 (3.19)	14.40	6.07	15.69
	5.5	13.54 (2.61)	14.35	7.35	16.73
	6.5	11.37 (3.55)	10.89	5.96	18.11
	7.5	8.94 (3.66)	9.00	3.69	13.16
Frequency of end-clause pauses	5.0	24.56 (4.27)	24.75	17.95	30.75
	5.5	24.38 (3.77)	23.59	17.92	30.87
	6.5	19.52 (4.18)	19.34	12.89	29.20
	7.5	19.65 (4.24)	18.83	12.18	27.46
Frequency of total repair	5.0	5.91 (3.30)	4.88	1.21	10.40
	5.5	4.63 (5.04)	3.01	0.00	17.06
	6.5	5.72 (2.49)	4.73	1.67	9.82
	7.5	5.19 (2.87)	5.40	1.14	10.35
	5.0	1.48	1.48	1.30	1.61

Bigram log frequency		(0.10)			
	5.5	1.54 (0.06)	1.57	1.42	1.62
	6.5	1.67 (0.03)	1.68	1.62	1.73
	7.5	1.78 (0.03)	1.79	1.74	1.83
Bigram proportion	5.0	0.50 (0.07)	0.51	0.34	0.58
	5.5	0.52 (0.06)	0.55	0.40	0.60
	6.5	0.52 (0.02)	0.52	0.46	0.55
	7.5	0.53 (0.00)	0.53	0.53	0.54
Bigram association MI	5.0	1.47 (0.20)	1.55	1.11	1.69
	5.5	1.47 (0.15)	1.38	1.29	1.78
	6.5	1.38 (0.05)	1.39	1.17	1.41
	7.5	1.36 (0.01)	1.36	1.34	1.38
Bigram association T	5.0	55.34 (25.57)	58.39	-7.76	83.93
	5.5	62.21 (19.86)	59.13	14.32	95.57
	6.5	62.41 (9.58)	65.01	26.61	67.90
	7.5	70.62 (1.51)	70.62	68.27	72.97
Trigram log frequency	5.0	0.83 (0.13)	0.85	0.58	0.98
	5.5	0.91 (0.17)	0.84	0.69	1.21
	6.5	0.96 (0.02)	0.96	0.93	0.99
	7.5	1.02 (0.02)	1.02	0.99	1.05
Trigram proportion	5.0	0.13 (0.05)	0.13	0.05	0.22
	5.5	0.16 (0.05)	0.17	0.09	0.25
	6.5	0.17 (0.01)	0.17	0.16	0.18
	7.5	0.19 (0.01)	0.19	0.18	0.19
Trigram association MI	5.0	2.24 (0.31)	2.22	1.86	2.76
	5.5	2.27	2.31	1.87	2.82

		(0.25)			
	6.5	1.38 (0.05)	2.21	1.83	2.22
	7.5	2.19 (0.01)	2.19	2.18	2.20
Trigram association T	5.0	19.59 (7.88)	20.53	7.88	30.79
	5.5	30.41 (12.96)	26.55	18.12	59.32
	6.5	30.61 (4.53)	31.50	13.92	33.96
	7.5	36.26 (1.29)	36.26	34.26	38.26

Use of MWSs at different proficiency levels

Our first research question was aimed at examining the use of MWSs at different proficiency levels. To examine the relationship between oral proficiency and n-gram measures, we conducted a Kruskal-Wallis H test, a non-parametric test which was adopted in response to the violation of homogeneity of variance across four proficiency groups (Levene statistic = 4.43 to 17.84, $p < .01$). The analysis showed that there were significant differences for six n-gram measures: Bigram frequency, $\chi^2(3) = 48.13$, $p = .001$; Bigram MI, $\chi^2(3) = 15.50$, $p = .001$; Bigram T, $\chi^2(3) = 12.54$, $p = .006$; Trigram frequency, $\chi^2(3) = 28.45$, $p = .001$; Trigram T, $\chi^2(3) = 31.18$, $p = .001$; Trigram proportion, $\chi^2(3) = 20.36$, $p = .001$, but not for two measures of Bigram proportion, $\chi^2(3) = 4.95$, $p = .175$ and Trigram MI, $\chi^2(3) = 7.15$, $p = .067$. Post-hoc tests (Mann-Whitney U tests) with Bonferroni correction revealed the following patterns at the $p < .008$ level: Bigram frequency (Low B1 = High B1 < B2 < C1); Bigram MI (High B1 = B2 > C1); Bigram T (B2 < C1); Trigram frequency (Low B1 < B2 < C1); Trigram T (Low B1 < B2 < C1); and Trigram proportion (Low B1 < B2 < C1). The analysis showed that for most measures, there was a linear relationship between general oral proficiency level and n-gram measures. Figures 1-8 provide a visual representation of the relationship between speaking proficiency and n-gram measures of frequency, proportion, and association (see Table 11 in supplementary materials for the information on effect sizes, confidence intervals, and *p* values).

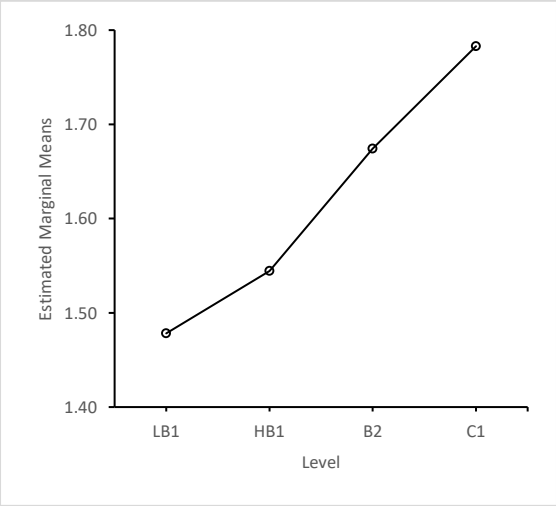

Figure 1: Bigram frequency & proficiency

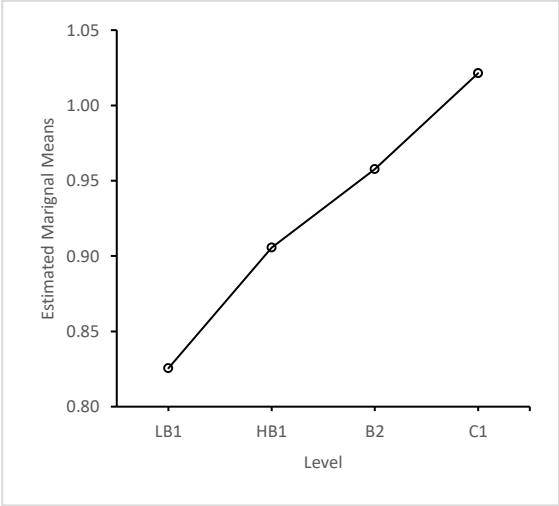

Figure 2: Trigram frequency & proficiency

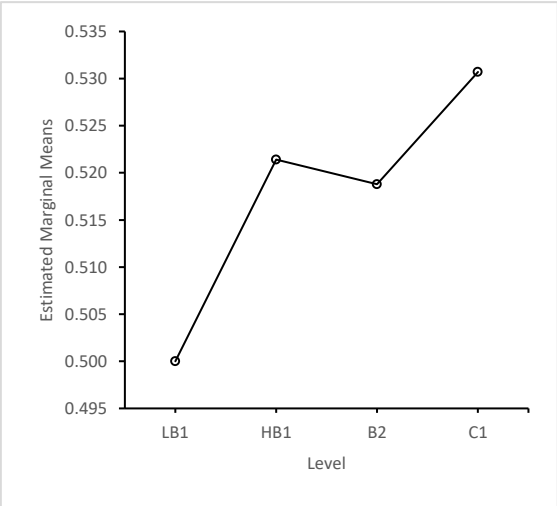

Figure 3: Bigram proportion & proficiency

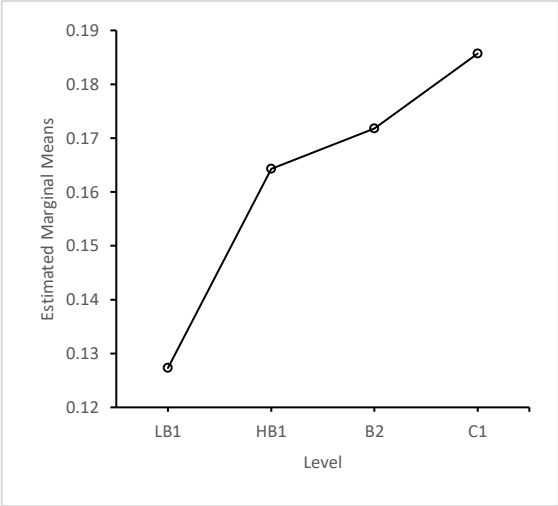

Figure 4: Trigram proportion & proficiency

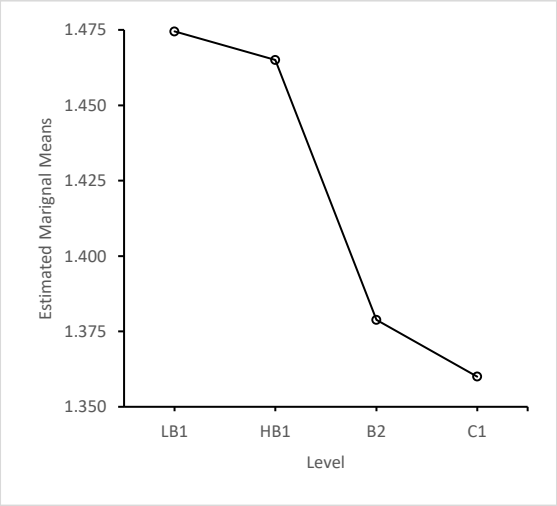

Figure 5: Bigram MI & proficiency

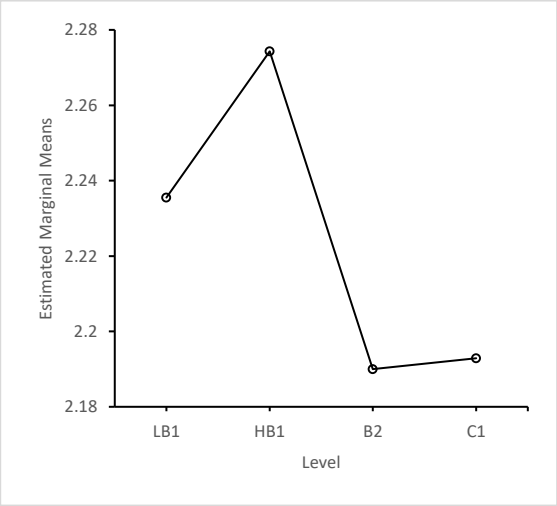

Figure 6: Trigram MI & proficiency

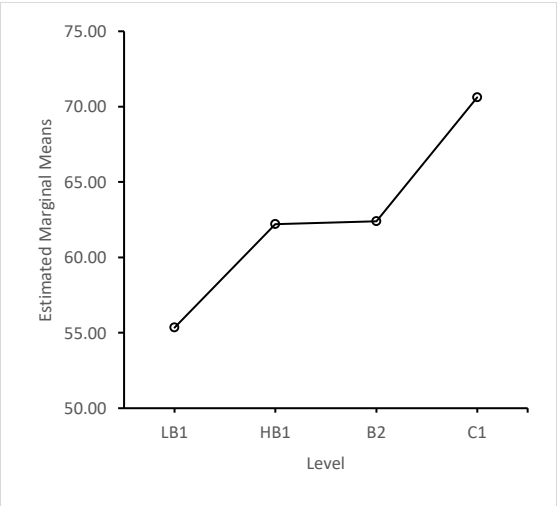

Figure 7: Bigram T-score & proficiency

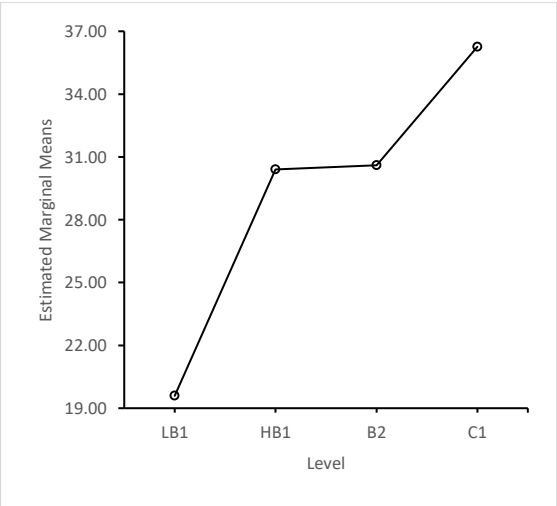

Figure 8: Trigram T-score & proficiency

Our second research question asked whether there was a relationship between different aspects of fluency (speed, breakdown, and repair) and different n-gram measures (proportion, frequency, and association). In total, we had four measures of fluency and eight measures of n-gram. In order to answer the question, we first ran a principal component analysis to see how many underlying factors formed the n-gram measures. Given the small sample size of the study (N = 56), we were concerned that running multiple correlations for all the n-gram measures would run a risk of Type 1 error. As our fluency measures were selected based on the findings of previous research (Suzuki & Kormos, 2019; Kahng, 2014), we believe they are valid representatives of the three aspects of fluency. In the next section, we discuss the results of the principal component analysis before reporting the correlation analyses.

Underlying factors of n-gram units

Data reduction. To control for the potential overlap between the MWSs measures, all eight n-gram measures were submitted to a principal component analysis with varimax rotation. The first attempt of the principal component analysis produced four factors, one of which was explained by a single n-gram measure alone (i.e., Bigram T-score). In order to avoid the issue of factor under-determination (i.e., factors are not represented by multiple measured variables) (Fabrigar et al., 1999), we conducted an additional principal component analysis using all n-gram measures excluding the Bigram T-score. The factorability of the entire dataset was confirmed using two tests: Bartlett's test of sphericity (χ^2 = 230.89, p = .001) and the Kaiser-Meyer-Olkin measure of sampling adequacy (.647). Based on eigenvalues beyond 0.7 (Stevens, 2002) and the visual inspection of the scree plot, three factors were identified in the data set, accounting for 86.1% of the variance in the n-gram measures (summarized in Table 3). We labelled Factor 1 "High Frequency Trigram" because it encompassed trigram measures that characterize high frequency combinations of words, that is, Trigram frequency and Association measure of T, which is reported to highlight high frequency combinations (Durrant & Schmitt, 2009). Although the result showed moderate sizes of loadings of Bigram frequency (r = .604) and Trigram proportion (r = .616) on Factor 1, we labelled this factor using two trigram measures of the higher loadings *(r* = .90).[ii] Factor 2 was labelled as "Association Measure of MI" because the result showed positive loadings of the two association measures (bigram and trigram MI scores) on this factor. The negative loading of bigram log frequency on this factor also looks reasonable given that MI scores are reported to highlight the low-frequency nature of the n-grams (Gablasova et al., 2017). Factor 3 was labelled "Proportion" as it captured bigram and trigram proportion scores. Given the small sample size of the study, we suggest the results of the factor analysis are considered cautiously. Table 3 below shows all the loadings for the three underlying factors.

Table 3 Principal component analysis of the n-gram measures

	Factor 1 (High Frequency Trigram)	Factor 2 (Association Measure of MI)	Factor 3 (Proportion)
Trigram log frequency	**.920**	-.089	.223
Trigram T-score	**.910**	.198	.125
Bigram MI	-.082	**.894**	.038
Trigram MI	.404	**.764**	-.278
Bigram log frequency	**.604**	**-.614**	.225
Bigram proportion (30,000)	.130	-.125	**.947**
Trigram proportion (30,000)	**.616**	-.062	**.711**

Relationship between fluency and n-gram units

Correlation analysis. To examine the relationship between fluency and n-gram units, a series of simple correlation analyses were conducted between the resulting three n-gram factors and fluency measures of speed, breakdown, and repair (Table 4). It is plausible to argue that unpruned data containing several repair features may not accurately represent speed fluency. To address this concern, we decided to include both pruned and unpruned measures of articulation rate to examine their relationship with the n-gram factors. Preliminary analyses were run to ensure there were no violations of the assumptions of normality, linearity, and homoscedasticity by visual inspection of histograms, scatterplots, P-P plots, and plots of studentized residuals crossed with fitted values (see Larson-Hall, 2010, p. 160). Significant but moderate correlations were observed between High Frequency Trigrams and pruned articulation rate (r = .396, p = .003*)* and unpruned articulation rate (r = .325, p = .014*)*, suggesting that fluent speakers tend to produce more high frequency n-grams. For frequency of mid-clause pauses, there was a negative correlation with n-gram Proportion factor (r = -.271, p = .041*)*, implying that speakers who paused more frequently in mid-clause positions tended to produce a lower proportion of target-like n-grams. There was a significant positive correlation between end-clause pauses and Association Measure of MI (r = .302, p = .024), suggesting that those who produced n-grams of high mutual information quality also paused more frequently at end-clause positions. For repair fluency, the MI measure was negatively correlated with frequency of total repair (r = -.308, p = .020*)*, indicating that speakers who made more repairs tended to produce more strongly associated n-grams.

Table 4 Pearson correlations between fluency and n-gram factor measures

		Correlations				
		Articulation Rate (unpruned)	Articulation Rate (pruned)	Frequency of mid-clause pauses	Frequency of end-clause pauses	Frequency of total repair
Factor 1 (High Frequency Trigram)	Pearson Correlation	.325*	.396**	-.239	-.202	-.073
	Significance (2-tailed)	.014	.003	.076	.136	.593
Factor 2 (Association Measure of MI)	Pearson Correlation	-.031	-.067	.215	.302*	-.308*
	Significance (2-tailed)	.821	.622	.112	.024	.021
Factor 3 (Proportion)	Pearson Correlation	.136	.165	-.271*	-.016	-.065
	Significance (2-tailed)	.316	.224	.041	.905	.633

Qualitative analysis of MWS use across different proficiency levels

To gain a deeper understanding of the role MWSs play in the development of oral proficiency, we examined the profiles of the most frequently used bigrams and trigrams across proficiency levels (Low B1, High B1, B2, and C1). Table 5 shows the 20 most frequent bigrams and trigrams used by all the participants (N = 56). Tables 5, 6, 7 and 8 (provided in the supplementary materials) indicate the breakdowns of the most frequent MWSs for Low B1 (n = 11), High B1 (n = 14), B2 (n = 17), and C1 (n = 14) groups respectively.

Table 5 The Most Frequent 20 Bigrams and Trigrams Used by All Speakers (*N* = 56)

		Text-internal		Text-external (COCA)				Text-internal		Text-external (COCA)		
	Bigram	**Freq**	**Range (N = 56)**	**Freq**	**MI**	**T**	**Trigram**	**Freq**	**Range (N = 56)**	**Freq**	**MI**	**T**
1	I think	138	39	2601.96	3.59	477.47	I think it	46	26	314.05	3.68	166.31
2	historical sites	104	32	n/a	n/a	n/a	a lot of	41	14	880.04	3.81	279.23
3	in the	74	34	4010.06	1.70	498.03	preserving historical sites	39	23	n/a	n/a	n/a
4	for the	73	34	1248.01	1.28	245.58	sites for leisure	35	16	n/a	n/a	n/a
5	and the	56	27	1593.61	0.30	100.44	building sites for	28	13	n/a	n/a	n/a
6	and it	54	29	801.70	0.77	146.11	I think the	22	11	232.46	3.75	143.34
7	of the	52	27	4667.36	1.57	521.34	good for the	21	12	10.22	1.73	25.32
8	it can	51	24	73.80	0.72	42.34	the historical sites	21	8	n/a	n/a	n/a
9	preserving historical	51	30	n/a	n/a	n/a	is good for	19	10	4.18	2.03	17.09
10	good for	50	19	44.13	1.39	48.08	historical sites is	18	10	n/a	n/a	n/a
11	it is	50	23	856.32	1.57	222.97	first of all	16	13	113.71	5.90	102.38
12	they can	50	23	186.17	2.34	118.67	for national identity	16	15	n/a	n/a	n/a
13	the country	48	26	308.93	2.21	150.71	important for national	16	15	n/a	n/a	n/a
14	for leisure	47	20	n/a	n/a	n/a	so I think	15	12	87.73	2.06	78.70
15	think it	47	27	440.90	2.02	175.41	in the past	14	7	75.09	3.43	80.71
16	lot of	46	17	916.72	3.56	283.18	of the country	13	10	52.87	2.08	61.29
17	a lot	45	16	1156.60	3.79	319.99	it's good for	13	6	n/a	n/a	n/a
18	for example	44	26	126.97	4.43	107.19	site for leisure	12	8	n/a	n/a	n/a
19	important for	41	27	21.08	1.93	37.81	to talk about	12	11	104.93	2.54	90.81
20	sites for	40	18	n/a	n/a	n/a	very important for	12	8	4.82	4.53	20.92

Note. Contracted forms (e.g., it's) were counted as a single word; for example, "it's good for" was considered to be a trigram rather than a quadgram. n/a indicates that the n-gram was not frequent enough in reference corpus for a score to be calculated.

As indicated in the tables, in general all the test-takers regardless of their proficiency level seem to have used similar types of bigrams and trigrams, including both grammatically complete (e.g., *historical sites*) and incomplete (e.g., *in the, and I think*) sequences. A closer inspection of the data, however, reveals some patterns emerging from the analysis. First, all the test-takers tend to use MWSs that are specifically relevant to task completion. Many of these MWSs, for example, *preserving historical sites, the historical sites, and sites for leisure,* seem to have been borrowed from the test/task instructions in which the test-takers were guided on what to discuss to complete the task. This observation is based on the test-takers' task performances (e.g., "As we know our topic is about X"). Compared to other speakers, those at C1 level seemed to show a greater tendency to borrow these n-grams, and to use them more frequently and competently.

As regards the frequent use of MWSs from task prompt, it should be noted that the novice speakers at a Low B1 level also used a large number of task-specific bigrams and trigrams (e.g., *historical sites, building sites for,* and *sites for leisure*). Despite this seemingly similar pattern between beginning (Low B1) and advanced (C1) learners regarding the use of task-related phrases, a closer inspection of the data indicated a qualitative difference in use of such MWSs between the two groups. While the Low B1 group borrowed the n-grams and used them repeatedly in their original form (e.g., *preserving historical sites*), the C1 level group borrowed the same n-grams but used them creatively by replacing some of the words with their synonyms (e.g., *protecting historical sites*), or by changing the structure of the n-grams (e.g., *to preserve these historical sites*). Overall, the lower proficiency speakers' speech was typically characterized by repetition and redundancy of the MWSs, whereas C1 level speakers avoided repetition and recycled the MWSs competently.

Finally, the qualitative inspection of the data highlighted an interesting pattern with respect to the n-gram use across different proficiency levels. The analysis revealed that the participants tended to use discourse markers (e.g., *for example, first of all*) or lexical fillers (e.g., *you know*) more frequently as proficiency increased. Table 6 below shows token counts for three most frequent discourse markers used by the participants. As shown in the table, an increase is proficiency is associated with an increase in the number of these discourse markers.

Table 6 Token counts of three phrases across four proficiency levels

	for example	*first of all*	*you know*
Low B1	1/11 (9%)	2/11 (18%)	2/11 (18%)
High B1	6/14 (43%)	2/14 (14%)	2/14 (14%)
B2	7/17 (41%)	4/17 (24%)	5/17 (29%)
C1	11/14 (79%)	6/14 (43%)	7/14 (50%)

Discussion

Answering the first research question regarding the relationship between use of MWSs and general oral proficiency, we found a linear relationship between many of the n-gram measures and proficiency level. This finding is important as it suggests that speakers at higher proficiency levels produce a greater proportion of frequent n-grams, more frequent n-grams, and n-grams with lower MI scores. While this finding is generally in line with previous research (Garner & Crossley, 2018; Kim et al., 2018; Kyle & Crossley, 2015), it provides significant additional evidence in support of the key role that MWSs play in oral proficiency. Given the scarcity of studies exploring use of n-grams in oral performance across a range of proficiency levels, and that proficiency was assessed by a standardized speaking test in our study, the results make a valuable contribution to the field of SLA.

The results confirm previous findings that higher proficiency learners produce a greater proportion of n-grams and a greater number of high frequency n-grams (Garner & Crossley, 2018; Kim et al.,

2018; Kyle & Crossley, 2015). These findings support the view that language proficiency is closely linked with the quality of MWSs (i.e., high frequency n-grams produced) as well as the size of MWSs (i.e., the number of n-grams produced). If we agree that language proficiency develops through exposure and practice, we may argue that proficient learners with more practice and exposure might be more sensitive to distributional patterns of language use than less proficient learners (Ellis, Simpson-Vlach, & Maynard, 2008). In this line of reasoning, high proficiency learners tend to use a large number of MWSs that are also frequent in the speech of L1 users.

However, the negative relationship between Bigram MI scores and oral proficiency observed in this study was not in line with previous research in either writing or speaking studies. The writing studies have documented that higher proficiency learners produce more strongly associated n-grams indexed by higher MI scores (Durrant & Schmitt, 2009; Granger & Bestgen, 2014), while speaking studies have reported little to no meaningful relationship between MI scores and oral proficiency (Garner & Crossley, 2018; Kim et al., 2018). In the current study, the results suggested that the highest proficiency group, compared to lower levels, did not produce a greater number of strongly associated n-grams (indexed by high MI scores). A possible reason for this unexpected finding might be linked to L2 beginners' repeated use of phrases with relatively high MI scores (e.g., MI ≧ 3; Hunston, 2002). A careful investigation of the transcripts of the data suggested that some of the n-grams, for example "I think" (MI = 3.59) were overused by learners of lower proficiency levels: 3.27 occurrences of the phrase per speaker for Low B1, 2.85 for High B1, 1.94 for B2, and 2.07 for C1. Furthermore, learners' production of strongly associated n-grams appeared to be greatly affected by their use of phrases borrowed from task prompts. The frequent use of task-specific phrases is particularly obvious in trigram use among C1 speakers (e.g., *preserving historical sites*), the majority of which occurred too infrequently (≦5 occurrences) in the COCA spoken sub-corpus for a score to be calculated (such atypical n-grams are marked with "n/a" in Table 5). This meant that quite a few of the n-grams produced by C1 speakers (e.g., *the historical sites, preserving historical sites, sites for leisure, building sites for, historical sites is*) did not occur frequently enough in a reference corpus to receive either frequency or association score computed by TAALES. Therefore, these n-grams were not considered in calculating the means of frequency or association scores per speaker. In fact, the correlation between frequency of n-grams used by C1 speakers and assigned MI scores is negligible (bigrams: r = .19, trigrams: r = .07), indicating that C1 speakers did not necessarily produce n-grams with higher MI scores repeatedly. By contrast, the correlation in the groups of lower-proficiency speakers is moderate as is observed for Low B1 (bigrams: r = .39) and High B1 (bigrams: r = .43, trigrams: r = .47). The negligible correlation for Low B1 trigrams (r = .07), can similarly be explained in terms of the participants' heavy reliance on task-specific key-phrases (e.g., *building sites for, sites for leisure, preserving historical sites*). It should be noted that since these MWSs did not occur frequently, they did not receive either frequency or association score in the n-gram analysis. Although our interpretation of the results is suggestive of task effects complicating the relationship between phraseological competence and L2 proficiency, our analysis is descriptive and speculative rather than conclusive. Future research can focus on this relationship to see if these results are replicated with different tasks (e.g., independent vs. integrated speaking tasks), and under different test conditions.

In answer to the second research question regarding the relationship between use of MWSs and oral fluency, moderate correlations were observed between several n-gram factors (High Frequency Trigram, Association Measure of MI, and Proportion) and fluency measures (articulation rate, frequency of mid-clause and end-clause pauses, frequency of total repair). First, the moderate positive correlation between speed fluency and high frequency n-grams suggested that fluent speakers with a high rate of speech produce more high frequency combinations of words (particularly more trigrams). This finding is in line with the psycholinguistic research evidence

suggesting that MWSs are stored and retrieved as individual units, allowing for information to be processed more quickly and conveniently (Ellis, 2012; Siyanova-Chanturia & Van Lancker Sidtis, 2018). We believe these findings can further help shed light on language processing in two ways, more specifically, one at the formulation stage and another at the articulation stage of speech production. First, at the formulation stage, learners with a large repertoire of MWSs can retrieve longer chunks of language holistically with similar processing load as required for retrieval of single-word items. With such retrieval efficiency, speakers might have a sufficient amount of remaining attentional resources which they can use for other aspects of language processing, such as syntactic and phonological encoding (Kormos, 2006). This processing efficiency might enhance speed fluency. Second, phraseological knowledge might also influence the later stage of speech production, that is, execution of articulation (Kormos, 2006) or, more specifically, articulation of individual words (Suzuki & Kormos, 2019). After phonemic segments or lexemes are mapped onto the selected lexical items (phonological encoding), speakers execute overt production drawing on such phonologically encoded information (Kormos, 2006; Levelt, 1989). This phonological information is hypothesized to serve as "addresses for stored phonetic syllable templates" or "motor instructions" for translating the phonologically filled frames into phonetic or articulatory program (Levelt, 1992, p. 16). Given that MWSs, particularly the high frequency ones, are subject to reduction of phonetic durations, for example, deletion of [t] in *I don't know* (Bybee & Scheibman, 1999), phraseologically proficient speakers could draw on articulatory instructions stored in their lexicon and produce utterances at a faster rate in general. Conversely, speakers lacking productive knowledge of MWSs may not have access to such information about phonetic reduction, and instead they may articulate every constituent word in full form (e.g., pronouncing the sound of [t] in *I don't know*), slowing down their overall articulation rate.

Our results also suggested that use of n-grams was related to breakdown fluency in terms of mid- and end-clause pausing. For end-clause pausing, we observed a positive correlation with MI measure, suggesting that when speakers used more strongly associated n-grams, they paused more frequently in end-clause positions. This is in line with previous research that suggests both L1 and L2 speakers are likely to pause before producing more difficult or sophisticated single-word items indexed by frequency (de Jong, 2016). The results also suggest that the speakers who used a larger proportion of MWSs paused less frequently at mid-clause positions. This is an important finding since mid-clause pausing is believed to be linked to the *formulation* stage of the speech production process (Skehan, 2014). SLA researchers (e.g., Kahng, 2014; Felkert, Clockman & de Jong, 2019) have long speculated that mid-clause pausing during L2 speech highlights the speakers' need to deal with the lexical and morphosyntactic demands of speech processing. Our findings support this view as they indicated that lack of phraseological knowledge increases the likelihood that learners pause in an unpredictable manner. Specifically, for less proficient learners there is a risk of retrieval failure in their attempt to access and select individual words, whereas phraseologically proficient learners are able to retrieve prefabricated chunks as a whole, providing them with "zones of safety" (Boers et al., 2006, p. 247), and therefore the risk of pausing would be confined to the spaces in between the MWSs in their utterance (e.g., syntactic or semantic boundary).

The results indicate that repair fluency is negatively linked with Association Measure of MI, indicating that speakers who demonstrated more strongly associated n-grams, repaired their speech less frequently. This finding implies that speakers producing MWSs of higher MI scores make fewer repetitions, reformulations, or self-corrections. It is also possible to argue that L2 learners who make more repairs draw less on formulaic language and target-like MWSs. It is essential to note that the existing literature has so far documented mixed findings concerning the role of repair phenomena in L2 proficiency and oral development (Saito et al., 2018; Tavakoli et al., in press). Some researchers, for example, Gilabert (2007), argue that repair behaviour is more closely linked with the accuracy rather than fluency aspect of performance due to L2 learners' attention to well-formedness.

The qualitative analysis of the use of n-grams across proficiency levels highlighted a few important patterns related to the use of n-grams. Firstly, the results underlined the important influence of task instructions on use of task specific MWSs, supporting the well-documented learner behaviour of text mining (i.e., transferring MWSs from input texts to L2 output; Boers et al., 2006; Hoang & Boers, 2016). The observed tendency of advanced speakers to use a great number of task-specific expressions accords with previous research exploring cohesion (or keyword and key-phrase overlap) between the task prompt and the speaker response and its impact on expert ratings of oral proficiency in the TOEFL-iBT integrated tasks (Crossley, Kyle, & Dascalu, 2019). Similar to the findings of the current study, Crossley et al. (2019) reported that advanced speakers used a number of key phrases (i.e., four-word sequences from the source text) in their speech, and that such speech samples were judged to be proficient by expert raters. Our analysis further revealed that advanced L2 speakers tried to avoid redundancy and repetition in their performance. Attempts to avoid redundancy in language use have been widely documented in previous corpus-based research (see Granger, 2018 for the review). Our results imply that attempts to avoid repetition and the ability to use MWSs competently should be considered a characteristic of advanced level L2 speakers. The qualitative analysis also highlighted an interesting pattern of use of discourse markers across proficiency level: The more proficient speakers used discourse markers more frequently. This finding supports earlier research documenting the increase in the use of discourse markers in oral production when proficiency develops (Tavakoli, 2018), and highlights the significant role of fillers as a useful communicative device for compensating for resource deficits in L2 processing and giving a positive impression about speakers (Préfontaine & Kormos, 2016).

Conclusions

The current study was aimed at providing a fine-grained picture of the role of MWSs in oral fluency across four levels of proficiency, assessed through a validated speaking test. To the best of our knowledge this is the first study that looks at this relationship in a systematic manner, both quantitatively and qualitatively, across four levels of proficiency. The findings confirmed that different aspects of oral fluency were associated with the proportion, frequency, and association strength of the MWSs produced by the speakers. More specifically, our results indicated that (a) greater use of high frequency MWSs in speech is linked to a faster articulation rate, (b) greater proportion of MWSs negatively correlates with the frequency of mid-clause pauses, (c) greater use of strongly associated MWSs negatively correlates with repair phenomena, and (d) greater use of strongly associated MWSs positively correlates with the frequency of end-clause pauses. Our qualitative analysis showed that some of the MWSs were borrowed from task instructions and were used frequently perhaps because they were seen as central to task completion. The findings also suggested that while low proficiency speakers relied on verbatim repetition of the borrowed MWSs, high proficiency speakers avoided repetition and instead used MWSs creatively. These findings are particularly important as they show how lexical demands of producing L2 speech relate to oral fluency, and that this relationship may vary at different levels of proficiency. The results are also important to language testing organizations as they can inform the development of rating descriptors and rating scales to evaluate formulaic use of language.

There are several limitations in the current study worth noting for future investigation of the relationship between MWSs and oral proficiency. First, the study used a relatively small sample size of 56 participants' data at only four levels of proficiency. Replicating these findings with a larger sample size and a wider spectrum of proficiency levels will provide a more in-depth insight into the relationship between MWS and oral proficiency. Second, the measurement of MWSs in this study was restricted to corpus-based measures, disregarding other linguistic features characterizing formulaicity, such as syntactic relations (e.g., adjective + noun, adverb + verb), semantic transparency (e.g., *kick a goal* vs. *kick the bucket*) and phonological features (e.g., intonation unit). Third, this study analyzed only two-word and three-word sequences in measuring MWSs, and

therefore the result may not be generalizable to longer sequences (e.g., four word and five word sequences). Future studies are needed to look into longer MWSs in general and their relationship to oral proficiency in particular. Fourth, we based our MWS measures on n-gram tokens rather than types. Future research should investigate the differences between analyses based on n-gram tokens versus n-gram types. Finally, like earlier lexical bundle studies (e.g., Huang, 2015), this study did not distinguish semantically incomplete n-grams (e.g., *the fact that)* from complete n-grams (e.g., *in other words*). A recent psycholinguistic study conducted by Jeong and Jiang (2019) found that semantically complete units were processed differently from incomplete units, suggesting the former is more psychologically valid than the latter as a holistically represented MWSs in the mental lexicon. Given this finding, future studies should consider semantic and syntactic completeness to explore the extent to which knowledge of MWSs predicts oral proficiency.

References

Ahmadian, M. J. (2011). The effect of 'massed' task repetitions on complexity, accuracy and fluency: Does it transfer to a new task? *The Language Learning Journal, 39*, 269-280. http://dx.doi.org/10.1080/09571736.2010.545239

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing, 26*, 18-41. http://dx.doi.org/10.1016/j.jslw.2014.09.004

Boers, F., Eyckmans, J., Kappel, K., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: putting a lexical approach to the test. *Language Teaching Research, 10*, 245-261. http://dx.doi.org/10.1191/1362168806lr195oa

Boers, F., & Webb, S. (2018). Teaching and learning collocation in adult second and foreign language learning. *Language Teaching, 51*, 77-89. http://dx.doi.org/10.1017/S0261444817000301

Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer (Version 5.3.39). Available: http://www.praat.org

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing, 30*, 159-175. http://dx.doi.org/10.1177/0265532212455394

Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: the reduction of don't in English. *Linguistics, 37*, 575-596. http://dx.doi.org/10.1515/ling.37.4.575

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

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods, 51*(1), 14-27. http://dx.doi.org/10.3758/s13428-018-1142-4

Davies, M. (2009). The 385+ million word *Corpus of Contemporary American English* (1990–2008+). Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*, 159–90

de Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching, 54,* 113-132. http://dx.doi.org/10.1515/iral-2016-9993

de Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In R. Eklund (Ed.), *Proceedings of the 6th workshop on disfluency in spontaneous speech (DiSS)* (pp. 17-20), Stockholm: Royal Institute of Technology (KTH).

de Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behaviour. *Applied Psycholinguistics, 36*, 223−243. http://dx.doi.org/10.1017/S0142716413000210

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. http://dx.doi.org/10.1515/iral.2009.007

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

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 42*(3), 375-396. http://dx.doi.org/10.1002/j.1545-7249.2008.tb00137.x

Evert, S. (2005). The statistics of word co-occurrences: Word pairs and collocations. Ph.D. thesis. Stuttgart: University of Stuttgart.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4,* 272-299. http://dx.doi.org/10.1037/1082-989X.4.3.272

Felker E. R., Klockmann H. E., & de Jong N. H. (2019). How conceptualizing influences fluency in first and second language speech production. *Applied Psycholinguistics, 40,* 111-136. http://dx.doi.org/10.1017/S0142716418000474

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21,* 354-75. http://dx.doi.org/10.1093/applin/21.3.354

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

Garner, J., & Crossley, S. (2018). A latent curve model approach to studying L2 N-Gram development. *The Modern Language Journal, 102*, 494-511. http://dx.doi.org/10.1111/modl.12494

Gilabert, R. (2007). Effects of manipulating task complexity on self repairs during L2 oral production. *International Review of Applied Linguistics in Language Teaching. 45*, 215–240. http://dx.doi.org/10.1515/iral.2007.010

Granger, S. (2018). Formulaic sequences in learner corpora. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 228–247). New York: Routledge.

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

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 27–49). Amsterdam: John Benjamins Publishing.

Hoang, H., & Boers, F. (2016). Re-telling a story in a second language: How well do adult learners mine an input text for multiword expressions? *Studies in Second Language Learning and Teaching*, *6*, 513–535. http://dx.doi.org/10.14746/ssllt.2016.6.3.7

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, *19*, 24–44. http://dx.doi.org/10.1093/applin/19.1.24

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. http://dx.doi.org/10.1016/j.system.2015.06.011

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

Jeong, H., & Jiang, N. (2019). Representation and processing of lexical bundles: Evidence from word monitoring. *System*, *80*, 188-198. http://dx.doi.org/10.1016/j.system.2018.11.009

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, *64*, 809-854. http://dx.doi.org/10.1111/lang.12084

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

Koponen, M., & Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In *Perspectives on fluency* (pp. 5−24). Ann Arbor, MI: University of Michigan.

Kormos, J. (1999). Monitoring and self-repair in L2. *Language Learning, 49*, 303-342. http://dx.doi.org/10.1111/0023-8333.00090

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*, 757-786. http://dx.doi.org/10.1002/tesq.194

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*, 1030-1046. http://dx.doi.org/10.3758/s13428-017-0924-4

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

Laufer, B., & Waldman, T. (2011). Verb-Noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*, 647-672. http://dx.doi.org/10.1111/j.1467-9922.2010.00621.x

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

Levelt, W. (1989). *Speaking: from intention to articulation*. Cambridge, MA: MIT Press.

Levelt, W. J. (1992). Accessing words in speech production: Stages, processes and representations. *Cognition, 42*, 1-22. http://dx.doi.org/10.1016/0010-0277(92)90038-J

Loewen, S. & Sato, M. (2018). State of the art article: Interaction and instructed second language acquisition. *Language Teaching*, *51*(3), 285-329. https://doi.org/10.1017/S0261444818000125

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly, 46*, 610-641. http://dx.doi.org/10.1002/tesq.34

Myles, F., & Cordier, C. (2017). Formulaic sequence (FS) cannot be an umbrella term in SLA. *Studies in Second Language Acquisition*, *39*, 3-28. http://dx.doi.org/10.1017/S027226311600036X

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*, 121-145. http://dx.doi.org/10.1177/0267658317694221

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

Paradis, M. (2009). *Declarative and Procedural Determinants of Second Languages*. Amsterdam: John Benjamins.

Pawley, A., & Syder, F. H. (1983). Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar. *Journal of Pragmatics, 7,* 551-579. http://dx.doi.org/10.1016/0378-2166(83)90081-4

Préfontaine, Y., & Kormos, J. (2016). A qualitative analysis of perceptions of fluency in second language French. *International Review of Applied Linguistics in Language Teaching, 54,* 151-169. http://dx.doi.org/10.1515/iral-2016-9995

Saito, K., Ilkan, M., Magne, V., Tran, M., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency. *Applied Psycholinguistics, 39,* 593-617. http://dx.doi.org/10.1017/S0142716417000571

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. London: Routledge.

Siyanova-Chanturia, A., & Martinez, R. (2015). The idiom principle revisited. *Applied Linguistics, 36,* 549-569. http://dx.doi.org/10.1093/applin/amt054

Siyanova-Chanturia, A., & Van Lancker Sidtis, D. (2018). What on-line processing tells us about formulaic language. In A. Siyanova-Chanturia and A. Pellicer-Sánchez (Eds.) *Understanding formulaic language: A second language acquisition perspective* (pp. 38-61). London, New York: Routledge.

Skehan, P. (1998). *A Cognitive Approach to language learning.* Oxford: Oxford University Press.

Skehan, P. (2003). Task-based instruction. *Language Teaching, 36,* 1-14. http://dx.doi.org/10.1017/S026144480200188X

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30,* 510-532. http://dx.doi.org/10.1093/applin/amp047

Skehan, P. (2014). *Processing perspectives on task performance*. Amsterdam: John Benjamins Publishing Company.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1,* 185–211. http://dx.doi.org/10.1177/136216889700100302

Skehan, P., Foster, P., & Shum, S. (2016). Ladders and snakes in second language fluency. *International Review of Applied Linguistics in Language Teaching, 54,* 97-111. http://dx.doi.org/10.1515/iral-2016-9992

Stengers, H., Boers, F., Housen, A., & Eyckmans, J. (2011). Formulaic sequences and L2 oral proficiency: does the type of target language influence the association? *International Review of Applied Linguistics, 49,* 321-343. http://dx.doi.org/10.1515/iral.2011.017

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Erlbaum.

Suzuki, S. & Kormos, J. (2019). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*. https://doi.org/10.1017/S0272263119000421

Suzuki, Y, & DeKeyser, R. (2017). Effects of distributed practice on the proceduralization of morphosyntax. *Language Teaching Research*, *38*, 27-56. http://dx.doi.org/10.1177/1362168815617334

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers, *ELT Journal. 65*, 71-79. http://dx.doi.org/10.1093/elt/ccq020

Tavakoli, P. (2018). L2 Development in an intensive Study Abroad EAP context. *System, 72*, 62-74. http://dx.doi.org/10.1016/j.system.2017.10.009

Tavakoli, P. & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.). *Planning and task performance in a second language* (pp. 239-277). Amsterdam: John Benjamins.

Tavakoli, P. Nakatsuhara, F. & Hunter, A-M. (2017). *Scoring validity of the Aptis Speaking test: Investigating fluency across tasks and levels of proficiency*. ARAGs Research Reports Online. ISSN 2057-5203London: British Council.

Tavakoli, P. Nakatsuhara, F. & Hunter, A-M. (in press). Aspects of fluency across assessed levels of speaking proficiency. *The Modern Language Journal*.

Thomson, H., Boers, F., & Coxhead, A. (2017). Replication research in pedagogical approaches to spoken fluency and formulaic sequences: A call for replication of Wood (2009) and Boers, Eyckmans, Kappel, Stengers & Demecheleer (2006). *Language Teaching, 52*, 1-9. http://dx.doi.org/10.1017/S0261444817000374

Wood, D. (2009). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *Canadian Journal of Applied Linguistics, 12*, 39-57.

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence, and classroom applications*. London, New York: Continuum.

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

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

Footnotes

[i] A third rater is invited to rate the samples independently if the first and second cannot agree.

[ii] We expected to observe strong and comparable loadings of bigram and trigram frequency indices on Frequency factor. However, since bigram and trigram frequency did not load onto the same factor, we decided to label Factor 1 as Trigram Frequency rather than Bigram and Trigram Frequency.