Studies in Second Language Acquisition (2023), **45**, 787–811
doi:10.1017/S0272263122000389

RESEARCH ARTICLE

Comparing the longitudinal development of phraseological complexity across oral and written tasks

Nathan Vandeweerd[1,2*], Alex Housen[2] and Magali Paquot[1]

[1]Université catholique de Louvain, Louvain-la-Neuve, Belgium; [2]Vrije Universiteit Brussel, Brussels, Belgium
*Corresponding author. E-mail: nathan.vandeweerd@uclouvain.be

(Received 19 October 2021; Revised 28 July 2022; Accepted 15 August 2022)

Abstract

This study builds upon previous research investigating the construct validity of phraseological complexity as an index of L2 development and proficiency. Whereas previous studies have focused on cross-sectional comparisons of written productions across proficiency levels, the current study compares the *longitudinal* development of phraseological complexity in written *and oral* productions elicited over a 21-month period from learners of French. We also improve upon the state of the art by including L1 data to benchmark learner levels of phraseological complexity. Phraseological complexity, operationalized as the diversity (no. types) and sophistication (PMI) of adjectival modifiers (adjective + noun) and direct objects (verb + noun), was generally higher in learner writing as compared to speaking. Over the study period, the sophistication of phraseological units increased slightly but developmental patterns were found to differ between tasks, highlighting the importance of considering task characteristics when measuring phraseological complexity.

Introduction

The use of complexity measures as indices of L2 development has a long history in second language research (see, e.g., Larsen-Freeman & Strom, 1977) and together with accuracy and fluency, complexity continues to be considered a key feature in the evaluation of L2 proficiency, that is to say, "a person's overall competence and ability to perform in L2" (Thomas, 1994, p. 330; Housen et al., 2012; Skehan, 1998). As discussed by Bulté and Housen (2014), the implicit assumption underlying complexity measures is that as a learner becomes more proficient, their linguistic output will incorporate more complex language and structures (e.g., a wider range of vocabulary, more infrequent lexical items, more sophisticated syntactic structures). Although there are a multitude of different ways that linguistic complexity can be operationalized, the most commonly used measures usually focus on solely lexical or syntactic aspects of a text (Bulté & Housen, 2012).

Paquot (2019) argues that while such measures are useful in their own right, they fail to fully capture the development of complexity in learners' interlanguage systems because they do not tap into complexity phenomena at the interface between lexis and grammar. As such, when used on their own, they have only limited validity as indices of L2 proficiency development. In particular, lexical and syntactic complexity measures fail to account for how words naturally combine to form conventional patterns of meaning and use (Sinclair, 1991).

The study of such word combinations is usually referred to under the umbrella term of *phraseology* and includes a wide variety of co-occurrence and recurrence phenomena including, for example, collocations, idioms, lexical bundles, colligations, and collostructions (Granger & Paquot, 2008). These units are phraseological in the sense that they involve "the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance" (Gries, 2008, p. 6). Such units have been shown to be an important component in the development of L2 proficiency. For example, various studies have found that compared to beginners, advanced learners use phraseological units with a higher pointwise mutual information[1] (PMI) (e.g., Granger & Bestgen, 2014; Kim et al., 2018; Kyle & Eguchi, 2021; Zhang & Li, 2021) and that the PMI of phraseological units used by learners increases over time (e.g., Bestgen & Granger, 2018; Edmonds & Gudmestad, 2021; Siyanova-Chanturia, 2015).

While the authors of these studies did not use the term *complexity*, Paquot (2019) argues that in measuring PMI, they were in fact tapping into the complexity of learners' phraseological systems. Paquot therefore attempted to build a bridge between L2 phraseology research and L2 complexity research by proposing the construct of *phraseological complexity*, which she defined as the diversity and sophistication of phraseological units (following Ortega, 2003). Phraseological diversity represents a learner's breadth of phraseological knowledge. In the same way that lexical diversity measures such as the type-token ratio index the number of different lexical units (words, lemmas, lexemes) that the learner uses, and are therefore indicative of the size of the learner's vocabulary, a learner production that displays a large number of different phraseological units is assumed to be indicative of a learner with a larger phraseological repertoire. Phraseological sophistication, on the other hand, represents the learner's knowledge of "sophisticated" phraseological units, those units that may be more specific and appropriate to a particular topic, register, or style. Thus, a learner production with greater phraseological sophistication would be indicative of a learner who is able to draw on a larger phraseological repertoire to select more specific or informative expressions.

Empirical evidence suggests that phraseological complexity develops with L2 proficiency. For example, using a corpus of linguistics term papers written by L2 learners of English, Paquot (2019) measured the complexity of three types of phraseological units that have been found to be difficult for L2 learners, namely, adjectival modifiers (e.g., black + hair), adverbial modifiers (e.g., eat + slowly), and direct objects (e.g., win + lottery). Each text in the corpus was manually assessed by a minimum of two

[1]Pointwise mutual information is a measure of the strength of association between words. When calculated on the basis of a large reference corpus, PMI has been found to highlight word combinations that are highly exclusive (Gablasova et al., 2017). Because of their exclusivity, phraseological units with a high PMI tend to be more topic specific, have more distinctive meanings, and tend to involve more specialized vocabulary.

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

professional raters who assigned it a global proficiency score ranging from B2 to C2 on the Common European Framework of Reference (CEFR) scale. When comparing across proficiency levels, Paquot found no significant differences in lexical or syntactic complexity but did find that phraseological sophistication, operationalized as the mean PMI of the units, increased significantly across proficiency levels. However, phraseological diversity, operationalized as the root type-token ratio (RTTR) of the units, was not significantly different across proficiency levels. Similar results were reported by Vandeweerd, Housen, and Paquot (2021) who conducted a partial replication study using L2 French argumentative essays (B2–C2). While the diversity of adjectival modifiers *did* increase significantly from C1 to C2, this measure was not found to be an important predictor of the scores given to a text by professional raters, when controlling for other aspects of complexity (i.e., lexical, syntactic, morphological).

These studies suggest that the construct of phraseological complexity (in particular sophistication)[2] can be a useful index of L2 proficiency and development at the upper levels of proficiency but they are limited by the fact that the focus has been exclusively on the written mode. To make generalizable claims about the construct validity of phraseological complexity, it is important to compare phraseological complexity measures across various tasks and performances (Purpura et al., 2015). Moreover, each of the previously mentioned studies has used a pseudo-longitudinal design, comparing phraseological complexity across learners at different proficiency levels. As Paquot and Granger (2012) point out, longitudinal studies are needed to identify patterns of phraseological development and allow the inference of more powerful cause and effect claims about L2 developmental processes (Ortega & Iberri-Shea, 2005). While there are a handful of studies that have compared the longitudinal development of phraseology (Bestgen & Granger, 2018; Edmonds & Gudmestad, 2021; Kim et al., 2018; Paquot et al., 2021; Qi & Ding, 2011), no study has yet directly compared the development of phraseology across oral and written L2 productions or compared the dimensions of diversity and sophistication across modes. This is important because not only the quantity but also the type of phraseological units may differ between modes (Biber et al., 2004) and studies have shown that the patterns of phraseological complexity observed in L2 writing do not necessarily carry over to L2 speech. For example, Paquot et al. (2022) measured the diversity (RTTR) and sophistication (PMI) of direct objects in a corpus of oral exams and found that, in contrast to the results for writing, there was a significant increase in diversity with proficiency (B2–C1) and a significant decrease in sophistication with proficiency (B1–B2), which they attributed to an increase in creativity on the part of the learners.

In comparison to writing, speech is usually more interactive, it tends to require more online processing and there is less control over output (Ravid & Tolchinsky, 2002). Importantly, because speech happens in real time, the limits of online processing constrain the amount of information that can be conveyed orally. This is especially evident in the case of L2 production due to the developing and less firmly entrenched state of the interlanguage system (De Bot, 1992). Under pressured conditions, writing is hypothesized to allow learners to devote more resources to conceptualizing, formulating, and monitoring linguistic output than is possible with speech because there is more opportunity for online (i.e., within task) planning (Skehan, 2014). Along these lines, several studies have found higher levels of lexical diversity and/or sophistication in

[2]The results of Rubin, Housen, and Paquot (2021) suggest that diversity measures may be more predictive at lower proficiency levels.

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

writing as compared to speaking in L2 production (e.g., Ellis & Yuan, 2005; Granfeldt, 2007; Kormos, 2014; Kuiken & Vedder, 2011). In a similar way, writing may also promote higher levels of phraseological complexity. Biber and Gray (2013) compared oral and written responses to L2 English proficiency exams by counting the number of collocations for five highly frequent verbs (*get*, *give*, *have*, *make*, and *take*). The results showed that spoken responses had a higher quantity of frequent collocates[3] for these verbs than written responses, suggesting that in the pressured situation of online speech production, learners were more likely to use highly frequent collocations but in the written task, learners were more likely to use less frequent and, hence, more sophisticated collocations. These results seem to be in line with psycholinguistic research showing that L2 learners process highly frequent collocations more quickly than infrequent collocations (Ellis et al., 2008). In other words, to the extent that writing allows for more online planning, it may promote the use of more sophisticated word combinations.

However, characterizing modality as a binary distinction between writing and speaking is somewhat reductive because modality often intersects with register. Studies of register variation in a number of languages have found that certain types of speech and writing have similar linguistic features because they serve a similar communicative purpose (Biber, 2019). For example, oral and written texts with a more "personal or involved" focus such as face-to-face conversations and personal letters tend to be more clausal in nature and are both characterized by the use of certain verb classes and similar types of verbal modification (Biber, 2014). Texts that are more informational in purpose, however, tend to be more nominal in nature and are characterized by the use of noun phrases, attributive adjectives, and more lexical diversity. The communicative function of a text is also relevant at the level of phraseology. In a study comparing spoken and written academic corpora, Biber, Conrad, and Cortes (2004) found that although oral corpora contained a higher quantity of lexical bundles overall compared to written corpora, conversation and classroom teaching differed with respect to the type of lexical bundles they contained. Compared to conversation, classroom teaching was associated with a higher proportion of noun phrase based "referential bundles," which the authors attributed to the informational purpose of those texts. These results suggest that both the diversity as well as sophistication of phraseological units used by a learner may be mediated by the register of a given task. For example, a learner may use a less diverse set of phraseological units because they are not required for the communicative purpose of a given task or they may use more sophisticated phraseological units because they are specifically elicited by the register characteristics of that task.

Regarding French, which is the target language of the learners in our study, no one has yet compared the development of phraseology across modes, but phraseological complexity has been shown to be associated with proficiency separately in each modality (Forsberg, 2010; Vandeweerd et al., 2021)[4] and has been shown to increase over time during a study abroad (Edmonds & Gudmestad, 2021). While these studies suggest that phraseology develops with proficiency in French, no study has yet traced the development of phraseological complexity in terms of diversity and sophistication over time, nor compared the longitudinal development across oral and written tasks.

[3]This was operationalized as words that appeared within a window of three words of the target verb in at least 10 texts at a rate of at least five times per 100,000 words.

[4]Forsberg (2010) did not use the term *phraseological complexity* but did measure the diversity (type-token ratio) and sophistication (PMI) of phraseological units in L2 French oral productions.

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

The current study aims to address these gaps by comparing phraseological complexity development in oral and written tasks completed by both learners of French as well as L1 users of French, thus also improving on the state of the art by following Housen and Kuiken's (2009) suggestion to include native benchmark data in studies investigating L2 complexity, accuracy, and fluency. Including L1 data in this way can help "explore which influences derive from tasks alone, rather than native speakerness" (Skehan & Foster, 2008, p. 204). However, it should be noted that the inclusion of native data is not meant to serve as a goal toward which the learners should strive (cf. recent criticisms of native-speakerism in L2 teaching and research by, among others, Holliday, 2006 and Ortega, 2019) but rather as a benchmark group whose cumulative experience as users of the target language means that their linguistic system is likely "richer, more accessible and better organized" (Skehan & Foster, 2008, p. 207) as compared to the learner group. In other words, L1 data can provide context for observed patterns of learner development (e.g., if the learners exhibit similar levels of phraseological complexity as the L1 group at the beginning of the study period, how much scope for further development can really be expected?). This question is especially important given that the construct of phraseological complexity is still new and so it remains an open question how much phraseological complexity is required to complete a task successfully and to date, no study of phraseological complexity has included a native benchmark group.

Our study sought to address these issues by investigating the complexity of two phraseological units (adjectival modifiers and direct objects) in a longitudinal corpus of oral and written production by learners of French and L1 users of French. Our research questions are the following:

RQ1. To what extent are there differences in phraseological complexity between oral and written tasks completed by the L1 group and the learner group?

RQ2. To what extent does the development of phraseological complexity differ between oral and written tasks completed by learners of French?

Methodology

Data

The data in this study come from the LANGSNAP corpus.[5] This corpus contains data from learners of French and Spanish in their second year of studies at a large university in the United Kingdom. The research team documented their language development over a 21-month period during which the students participated in a 9-month sojourn abroad. Data were also collected from a comparable group of L1 speakers (n = 10) who were Erasmus exchange students at the UK university. In the current study, we focus on the 29 learners of French as an additional language and the 10 native French speakers. The majority of the learners are L1 English speakers (n = 27) but the group also includes one L1 Spanish speaker and one L1 Finnish speaker. The mean age of the learner group is 21 (range 20–24) and includes 26 females and three males.[6] Prior to participation in the project, the learners had on average 11 years (range 9–15) of previous studies in French in an institutional setting. In addition to

[5]http://langsnap.soton.ac.uk/

[6]According to the compilers of the LANGSNAP corpus, this gender imbalance is representative of language programs in this setting (Mitchell et al., 2017, p. 52).

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

Table 1. Data collection schedule (adapted from Mitchell, Tracy-Ventura, and McManus, 2017, p. 56)

Data Collection Cycle	Location	Oral Tasks	Written Essay Topic
Presojourn (May 2011)	Home	Oral Interview Cat Story	Gay Marriage and Adoption
Insojourn (Oct. 2011)	Abroad	Oral Interview Sisters Story	Legalization of Marijuana
Insojourn (Feb. 2012)	Abroad	Oral Interview Brothers Story	Taxes on Junk Food
Insojourn (May 2012)	Abroad	Oral Interview Cat Story	Gay Marriage and Adoption
Postsojourn (Oct. 2012)	Home	Oral Interview Sisters Story	Legalization of Marijuana
Postsojourn (Feb. 2013)	Home	Oral Interview Brothers Story	Taxes on Junk Food

French, some of the participants were also studying other European languages (e.g., German, Spanish, Portuguese). For this reason, we refer to them as "learners of French" throughout. The background of the participants as well as their individual trajectories over the course of their stay abroad has been extensively reported elsewhere (see, e.g., Mitchell et al., 2017) so we will not go into further detail here.

Data were collected from the learners on six occasions before, during, and following their stay abroad in France (see Table 1). At each collection point, they completed three production tasks. As a measure of general proficiency, the participants also completed an Elicited Imitation Test (EIT) at three time points (for details see Tracy-Ventura et al., 2014). The three production tasks included a written argumentative essay, a semi-guided oral interview and a picture-based oral narrative. The argumentative essay task was set up to run offline on stand-alone computers. Participants were provided with one of three essay prompts in the following list, each of which was repeated once during the study:

- *Penses-tu que les couples homosexuels ont le droit de se marier et d'adopter des enfants ?*
  'Do you think that homosexual couples have the right to get married and adopt children?'
- *Pensez-vous que la marijuana devrait être légalisée ?*
  'Do you think that marijuana should be legalized?'
- *Pensez-vous que, de manière à inciter les gens à manger sainement, on devrait taxer les boissons sucrées et les aliments gras ?*
  'Do you think that in order to encourage people to eat in a healthy manner, sugary beverages and fatty foods should be taxed?'

Participants were allowed three minutes for planning and note taking before being automatically taken to the writing page where they had 15 minutes to write approximately 200 words. After the time had elapsed, the program exited automatically and the participants could no longer edit the text. The oral narrative task involved an oral description of one of three picture-based narratives. Participants were given a short amount of planning time, during which they could ask clarification questions. Each of the stories were set in the past and contained a similar number of color images. They were also designed to elicit perfective and habitual events and contained two characters. The oral interview was a semi-structured interview administered by one of the researchers. Each interview followed a list of preestablished questions about the participants' anticipations and experiences during their sojourn abroad. Questions were designed to elicit discussion of present, future, past, and hypothetical events.

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

Table 2. Median text lengths across tasks at each collection point (IQR in brackets)

Cycle	Argumentative Essay	Oral Interview	Oral Narrative
Presojourn 1	210 (30)	1,073 (475)	344 (120)
Insojourn 1	205 (24)	1,630 (955)	409 (143.5)
Insojourn 2	209 (23)	1,476 (888)	277 (112)
Insojourn 3	216.5 (32.25)	1,105 (786.5)	324.5 (116.75)
Postsojourn 1	211 (30)	692 (264)	402 (69)
Postsojourn 2	217 (12)	812 (370)	253 (92)

To facilitate the extraction of phraseological units, the original CHAT transcripts of these tasks were converted to plain text files using in house R scripts (R Core Team, 2021). The text files were then lemmatized and part of speech tagged using TreeTagger (Schmid, 1994). Table 2 shows the median text lengths for each of the task types following preprocessing.

Phraseological Complexity Measures

We follow Gries's (2008) definition of phraseological units as the co-occurrence of two linguistic units at a frequency higher than expected on the basis of chance. Because phraseological complexity research is still relatively new, only a handful of possible phraseological units have thus far been examined. Two units, namely adjectival modifier (adjective + noun) and direct object (verb + noun) relations have received particular attention because previous research has shown that L2 productions tend to exhibit less variety in these units and that learners tend to produce units with lower collocational strength as compared to native speakers (Altenberg & Granger, 2001; Durrant & Schmitt, 2009; Laufer & Waldman, 2011; Nesselhauf, 2005). We have decided to focus on these two units for the purpose of this study as well, given that the complexity of these units has been found to increase with proficiency levels in previous studies of phraseological complexity in L2 English, L2 Dutch, and L2 French (Paquot, 2019; Rubin et al., 2021; Vandeweerd et al., 2021). The inclusion of one noun phrase-based unit and one verb phrase-based unit also taps into the potential register differences that may be induced by the different task types (as shown by Biber et al., 2004). Specifically, in line with the literature on register variation, we expect the more personal or involved focus of the narrative and interview tasks will promote the complexity of direct objects whereas the informational focus of the essay will instead promote the complexity of in adjectival modifiers.

We wrote an R script to search the lemmatized versions of the texts and extract adjectival modifiers and direct objects using the part of speech tags generated by TreeTagger. The decision to use the lemmatized version of texts is due to the highly inflected nature of French (see Treffers-Daller, 2013). This is particularly important when comparing speech and writing given that many verbal inflections are only marked orthographically (and not phonologically) in French (see Blanche-Benveniste & Adam, 1999). To evaluate the reliability of the extraction method, a subset of 100 sentences from the written essays and 100 utterances[7] from the oral tasks was annotated by two researchers for the presence of adjectival modifier and

[7]The oral data was segmented into utterances following the CHILDES conventions (see MacWhinney, 2000). These utterances are henceforth referred to as *sentences*.

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

Table 3. Comparison of manual and automatic annotation

	Adjectival Modifiers		Direct Objects	
	Oral	Written	Oral	Written
Manual (n)	115	415	211	490
Automatic (n)	117	389	222	487
Precision	0.95	0.90	0.82	0.91
Recall	0.96	0.84	0.86	0.90
F-score	0.96	0.87	0.84	0.90

direct object relations following the definitions in Appendix 1. The annotators reached a high level of agreement for the units in both written ($\kappa_{amod} = 0.92$, $\kappa_{dobj} = 0.84$) and oral data ($\kappa_{amod} = 0.94$, $\kappa_{dobj} = 0.95$). All kappa values are above the minimal thresholds for reliability in SLA (0.83) according to Plonsky and Derrick (2016). In total, there were 39 cases of disagreement between the two annotators. Each of these cases were discussed and the annotation guidelines were refined. Following these discussions, one researcher annotated a further 400 oral and 400 written sentences. The combined sets of 500 written and oral sentences were then used as baseline to evaluate the reliability of the automatic method. As shown in Table 3, F-scores, which represent the balance between precision (not identifying incorrect units) and recall (not failing to identify correct units) were found to be acceptable for both types of units. Table 4 shows the median number of adjectival modifiers and direct objects tokens per 100 words across the three task types.

Following Paquot (2019), we measured phraseological complexity in terms of diversity and sophistication. Because of the large differences between the tasks regarding text length (see Table 2), measures were not calculated based on the entire text but rather as the average over 100-word[8] moving windows, moving the window forward by an increment of 10 words after each sample. This is similar to a commonly used method used to control for text-length differences when calculating lexical diversity (MATTR; Covington & McFall, 2010). A moving window-based approach was necessary because transformations of type-token ratios such as Guiraud's (1954) root type-token ratio (as used by Paquot, 2019 and Vandeweerd et al., 2021) were found to be correlated with text length. Other more sophisticated diversity measures such as MTLD (McCarthy & Jarvis, 2010) also could not be used because they require a minimum number of units (at least 100 as suggested by Koizumi, 2012) and some texts in the corpus had few to no phraseological units of a given type.[9] To maintain comparability between the diversity

Table 4. Median number of tokens per 100 words (IQR)

Task	Adjectival Modifiers	Direct Objects
Argumentative Essay	3.18 (2.1)	4.36 (1.49)
Oral Narrative	1.18 (0.92)	4.14 (1.49)
Oral Interview	1.96 (0.78)	3.62 (0.94)

[8]The window size of 100 words was chosen because the shortest text in the corpus was 110 words.

[9]As pointed out by one anonymous reviewer, this *could* hypothetically give greater weight to units occurring at the beginning of texts than to units occurring at the end of texts. In response to this comment, we analyzed the distribution of units throughout the texts and found very low correlations ($\tau < 0.01$) between position in a text and the number of units, suggesting that the units were fairly evenly distributed throughout texts.

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

and sophistication measures, we used this method for both types of measures. Phraseological diversity was operationalized as the average number of unique adjectival modifier and direct object types (i.e., unique units) in 100-word windows. Phraseological sophistication was operationalized as the mean PMI[10] score of adjectival modifiers and direct objects in 100-word windows.[11] As in Paquot (2021), PMI was calculated on the basis of a large web-scraped reference corpus. The reference corpus used in this study was the 10-billion-word FRCOW16 corpus (Schäfer, 2015; Schäfer & Bildhauer, 2012). The FRCOW16 corpus is provided with dependency annotation generated by Malt Parser (Nivre et al., 2006) trained on the French Tree Bank (Abeillé & Barrier, 2004). Adjectival modifiers and direct objects were extracted from the corpus on the basis of these dependency annotations using R scripts. We did not directly test the accuracy of this extraction method for the reference corpus given that it has previously been shown to have a relatively high level of accuracy (87% labeled attachment; Candito et al., 2010). When it comes to adjectival modifiers and direct objects, we also reported high levels of reliability (F-scores of 0.81 and 0.82, respectively) for these units in a previous study (Vandeweerd et al., 2021). In addition, the FRCOW16 corpus contains POS tags generated by TreeTagger, the same POS tagger used to process the learner corpus. Following the method described by Paquot (2019), a PMI value was calculated on the basis of the FRCOW16 corpus for the lemmatized version of each of the phraseological units extracted from the learner corpus. Units that contained a lemma unknown to TreeTagger (e.g., proper nouns, calques from English) were excluded as were units that occurred fewer than five times in the reference corpus and units that occurred in the writing prompts or were provided in the picture stories. The final list of extracted units was also checked and obvious examples of erroneous POS tags were also removed (e.g., *hier*, "yesterday" tagged as a verb). The mean PMI for adjectival modifiers and direct objects within each 100-word moving window was calculated for every text.

Analysis

We built mixed-effects regression models using the *nlme* package (Pinheiro et al., 2020) in R to predict each of the four phraseological complexity variables as outlined in the preceding text: (a) adjectival modifier types, (b) adjectival modifier PMI, (c) direct object types, and (d) direct object PMI. In each model, the phraseological complexity variable was the dependent variable. The main independent variables in the model were time (in months), task type, and the interaction between time and task type. Separate models were run for the L1 and learner data. For the learner models, given that previous study abroad research has shown a positive effect of initial proficiency on development while abroad (DeKeyser, 2007), we additionally included an interaction effect between EIT.PRE (initial proficiency) and time in months to control for the moderating effect of initial proficiency. To control for topic differences (both between prompts and between tasks), we calculated the cosine similarity between each text and the first argumentative essay written by the same learner. This measure represents the similarity of vocabulary between the two texts and ranges from 0 (indicating no overlap in vocabulary) to

[10] $PMI = log_2 \frac{P(x, y)}{P(x)P(y)}$ where $p(x, y)$ represents the probability of encountering x and y together and $p(x)$, $p(y)$ represent the probability of encountering x and y separately (Church & Hanks, 1990, p. 23).

[11] Eleven texts in the corpus contained no adjectival modifiers and so no PMI values could be assigned. These texts were excluded from the analysis of adjectival modifier sophistication.

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

1 (Wang & Dong, 2020). Including cosine similarity in the model allows us to account for the variation in phraseological complexity that is due to differences in vocabulary between different prompts and task types. Because this measure is quite right-skewed (most texts are dissimilar from each other), we applied a log transformation. Prior to modeling, the variables EIT.PRE (presojourn EIT scores) and COSINE.log (log of cosine similarity) were converted to z-scores. The random effects structure included random intercepts and by-participant random slopes for months. Planned orthogonal contrasts were set to compare written versus oral tasks as a whole and narrative versus interviews. The L1 models included only one fixed effect (task type) and included random intercepts for participants given that each participant provided seven different productions (for each task type and each topic). Initial models revealed problems with homoscedasticity that were resolved by weighting variance according to task type. For maximum comparability, we included all predictors in each model instead of using a model selection approach. This allows us to make clearer comparisons between models regarding our principle research questions.

The supplementary materials for this article are provided on an OSF repository.[12] These include the scripts used to preprocess the texts and extract phraseological units from the two corpora, further details about the process used to verify the reliability of the extraction method and calculate cosine reliability as well as the full model fitting procedure.

Results

The Learner Models

As shown in Table 5, significant predictors of adjectival modifier diversity (number of types) in the learner data included task type (both written vs. oral and narrative vs. interview), the interaction between task type (narrative vs. interview) and months

Table 5. Adjectival modifier diversity (learner data)

AMOD.TYPES ~ COSINE.log + (EIT.PRE * MONTHS) + (TASKTYPE * MONTHS)						
Fixed Effects	b	95% CI	SE	df	t	p
(Intercept)	2.162	[1.903, 2.421]	0.132	477	16.350	<0.001
COSINE.log	0.075	[0.008, 0.142]	0.034	477	2.182	0.030
EIT.PRE	−0.114	[−0.226, −0.001]	0.057	27	−1.981	0.058
MONTHS	0.003	[−0.011, 0.016]	0.007	477	0.357	0.721
TASKTYPEwritten→oral	−0.873	[−1.177, −0.570]	0.155	477	−5.638	<0.001
TASKTYPEnar→int	0.350	[0.234, 0.465]	0.059	477	5.944	<0.001
EIT.PRE:MONTHS	0.006	[−0.002, 0.015]	0.004	477	1.417	0.157
MONTHS:TASKTYPEwritten→oral	−0.012	[−0.037, 0.013]	0.013	477	−0.905	0.366
MONTHS:TASKTYPEnar→int	−0.009	[−0.018, 0.000]	0.005	477	−1.966	0.050
Random Effects	variance	sd				
(Intercept)	0.024	0.155				
MONTHS	<0.001	<0.001				
Residual	0.354	0.595				

Marginal R^2: 0.631
Conditional R^2: 0.655

[12] https://osf.io/bkfru/

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

Table 6. Linear trends for adjectival modifier types (learner data)

			vs. Narrative		vs. Interview	
Task	b	95% CI	t	p	t	p
Essay	0.014	[–0.023, 0.051]	0.420	0.907	1.355	0.366
Narrative	0.006	[–0.008, 0.020]			1.966	0.122
Interview	–0.012	[–0.023, -0.001]				

(though this is right on the threshold for significance at an alpha level of 5%), and cosine similarity. In general, the written argumentative essays were found to have about 0.87 more adjectival modifier types per 100 words than both of the oral tasks and the interview task was also found to have 0.35 more adjectival modifier types than the narrative task. In terms of developmental trends, there was a significant effect of time in months, but this was not the same across all task types given that there was a significant interaction effect of task type and time in months. As shown in in Table 6, development was only significant in the case of the interviews, which decreased by 0.01 per month (the 95% confidence intervals for the slope (b) of the essay and the narrative overlap with 0). No other significant effects of time were found for the other task types and neither initial proficiency nor the interaction of initial proficiency and time in months was significant, indicating that initial proficiency did not have a direct effect on the number of adjectival modifier types used nor on the development of adjectival modifier types over the study period. There was, however, a significant effect of topic (COSINE.log) such that a one standard deviation increase in vocabulary similarity to the gay marriage and adoption essay was associated with the use of 0.07 more adjectival modifier types. In total, 65.47% of the variance in the number of adjectival modifier types was explained by this model.

As shown in Table 7, the only significant predictor of adjectival modifier sophistication (PMI) was task type (written vs. oral). The average PMI of adjectival modifiers in written argumentative essays was higher by 0.38 as compared to the oral tasks. Neither the main effect of time in months nor the interaction of time in months with task type or

Table 7. Adjectival modifier sophistication (learner data)

AMOD.MEANPMI ~ COSINE.log + (EIT.PRE * MONTHS) + (TASKTYPE * MONTHS)						
Fixed Effects	b	95% CI	SE	df	t	p
(Intercept)	1.097	[0.803, 1.391]	0.150	468	7.306	<0.001
COSINE.log	–0.027	[–0.108, 0.054]	0.041	468	–0.656	0.512
EIT.PRE	0.137	[0.001, 0.273]	0.069	27	1.971	0.059
MONTHS	0.007	[–0.007, 0.022]	0.007	468	0.992	0.321
TASKTYPEwritten→oral	–0.378	[–0.634, –0.122]	0.131	468	–2.892	0.004
TASKTYPEnar→int	0.092	[–0.116, 0.300]	0.106	468	0.871	0.384
EIT.PRE:MONTHS	–0.004	[–0.015, 0.007]	0.006	468	–0.732	0.464
MONTHS:TASKTYPEwritten→oral	–0.010	[–0.030, 0.009]	0.010	468	–1.019	0.309
MONTHS:TASKTYPEnar→int	–0.011	[–0.028, 0.006]	0.009	468	–1.298	0.195
Random Effects	variance	sd				
(Intercept)	0.022	0.150				
MONTHS	<0.001	<0.001				
Residual	1.685	1.298				

Marginal R^2: 0.067
Conditional R^2: 0.08

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

Table 8. Direct object diversity (learner data)

DOBJ.TYPES ~ COSINE.log + (EIT.PRE * MONTHS) + (TASKTYPE * MONTHS)						
Fixed Effects	b	95% CI	SE	df	t	p
(Intercept)	3.640	[3.333, 3.948]	0.157	477	23.206	<0.001
COSINE.log	−0.030	[−0.115, 0.054]	0.043	477	−0.709	0.479
EIT.PRE	−0.030	[−0.184, 0.125]	0.079	27	−0.374	0.711
MONTHS	−0.007	[−0.021, 0.008]	0.007	477	−0.911	0.363
TASKTYPEwritten→oral	−0.378	[−0.666, −0.09]	0.147	477	−2.573	0.010
TASKTYPEnar→int	−0.239	[−0.397, −0.081]	0.080	477	−2.966	0.003
EIT.PRE:MONTHS	0.006	[−0.005, 0.017]	0.006	477	1.132	0.258
MONTHS:TASKTYPEwritten→oral	−0.016	[−0.039, 0.007]	0.012	477	−1.390	0.165
MONTHS:TASKTYPEnar→int	−0.007	[−0.019, 0.005]	0.006	477	−1.097	0.273
Random Effects	variance	sd				
(Intercept)	0.063	0.251				
MONTHS	<0.001	<0.001				
Residual	0.770	0.878				

Marginal R^2: 0.201
Conditional R^2: 0.262

initial proficiency was significant. The effect of topic was likewise nonsignificant. The predictors in this model explained 7.97% of the variance in adjectival modifier PMI.

As shown in Table 8, the only significant predictor of direct object diversity (number of types) was task type (both written vs. oral and narrative vs. interview). In general, the written argumentative essays were found to have about 0.38 more direct object types per 100 words than both of the oral tasks and the narrative task was found to have 0.24 more direct object types than the interview task. Neither the main effect of time in months nor the interaction of time in months with task type or initial proficiency was significant. The effect of topic was also found to be nonsignificant. This model explained 26.16% of the variance in direct object types.

As shown in Table 9, significant predictors of direct object sophistication (PMI) included task type (narrative vs. interview), time in months, and cosine similarity. The average PMI of direct objects in narratives was found to be higher by 0.12 as compared to interviews. In terms of developmental trends, the sophistication of direct objects increased by 0.01 per month. This developmental rate did not differ significantly across tasks (the interactions between task type and months were n.s.). Neither initial proficiency nor the interaction of initial proficiency and time in months was significant. There was, however, a significant effect of topic such that a one standard deviation increase in vocabulary similarity to the gay marriage and adoption essay was associated with decrease in direct object PMI of 0.06. This model explained 14.4% of the variance in direct object PMI.

The L1 Models

We have limited the discussion here of the L1 models to the estimated marginal means and pairwise comparisons for each variable according to task type but the full model output and diagnostics are available in the supplementary materials. As shown in Table 10, adjectival modifier diversity (number of types) was significantly higher in the argumentative essays (Mean = 3.47, CI = 2.89, 4.04) as compared to the narratives

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

Table 9. Direct object sophistication (learner data)

DOBJ.MEANPMI ~ COSINE.log + (EIT.PRE * MONTHS) + (TASKTYPE * MONTHS)						
Fixed Effects	b	95% CI	SE	df	t	p
(Intercept)	0.766	[0.598, 0.934]	0.086	477	8.956	<0.001
COSINE.log	−0.057	[−0.105, −0.010]	0.024	477	−2.353	0.019
EIT.PRE	0.055	[−0.020, 0.129]	0.038	27	1.434	0.163
MONTHS	0.011	[0.003, 0.019]	0.004	477	2.607	0.009
TASKTYPEwritten→oral	0.072	[−0.093, 0.237]	0.084	477	0.856	0.393
TASKTYPEnar→int	−0.116	[−0.210, −0.023]	0.048	477	−2.433	0.015
EIT.PRE:MONTHS	−0.002	[−0.009, 0.004]	0.003	477	−0.733	0.464
MONTHS:TASKTYPEwritten→oral	−0.007	[−0.020, 0.006]	0.007	477	−1.105	0.270
MONTHS:TASKTYPEnar→int	−0.004	[−0.011, 0.003]	0.004	477	−1.050	0.294
Random Effects	variance	sd				
(Intercept)	<0.001	0.021				
MONTHS	<0.001	0.004				
Residual	0.275	0.524				

Marginal R^2: 0.136
Conditional R^2: 0.144

(Mean = 1.12, CI = 0.78, 1.46) and the interviews (Mean = 1.44, CI = 1.16, 1.73) and adjectival modifier diversity between the two oral tasks was not significantly different. A different pattern was found for the three other phraseological complexity measures. The sophistication of adjectival modifiers (PMI) was significantly higher in both the essays (Mean = 2.15, CI = 1.80, 2.50) and the narratives (Mean = 1.93, CI = 1.54, 2.33) than in the interviews (Mean = 1.20, CI = 0.83, 1.57) and the difference between the essays and the narratives was not found to be significant. Similarly, direct object diversity (number of types) was found to be significantly higher in both the essays (Mean = 4.21, CI = 3.53, 4.89) and the narratives (Mean = 3.86, CI = 3.44, 4.29) than in the interviews (Mean = 2.98, CI = 2.76, 3.20), and the difference between the essays and the narratives was not found to be significant. Lastly, direct object sophistication (PMI) was significantly higher in the narratives (Mean = 1.55, CI = 1.39, 1.72) as compared to the interviews (Mean = 1.09, CI = 0.83, 1.35). The mean PMI of direct objects was also higher in the essays (Mean = 1.54, CI = 1.19, 1.90) as compared to the interviews but this difference was just beyond the threshold for significance at an alpha level of 5%. Again, the difference between the essays and the narratives was not significant. To summarize, written argumentative essays were found to contain a higher number of adjectival modifier types as compared to the oral tasks. However, for the three other measures, namely the number of direct object types as well as the mean PMI of both adjectival modifier and direct object relations, higher values were found in the written essay and oral narrative as compared to the oral interview task.

Discussion

RQ1. To what extent are there differences in phraseological complexity between oral and written tasks completed by the L1 group and the learner group?

In general, the written learner productions exhibited higher levels of phraseological complexity as compared to the oral productions. Specifically, the argumentative essays were found to use more adjectival modifier and direct object types and used adjectival

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

Table 10. Task differences for L1 Group (n = 10)

	Estimated Marginal Means [95% CI]			T-ratio (p)		
Measure	Essay	Narrative	Interview	Essay vs. Narrative	Essay vs. Interview	Narrative vs. Interview
AMOD.TYPES	3.47 [2.89, 4.04]	1.12 [0.78, 1.46]	1.44 [1.16, 1.73]	8.41 (<0.001)	7.6 (<0.001)	−1.89 (0.150)
AMOD.MEANPMI	2.15 [1.80, 2.50]	1.93 [1.54, 2.33]	1.20 [0.83, 1.57]	0.93 (0.621)	4.22 (<0.001)	3.07 (0.009)
DOBJ.TYPES	4.21 [3.53, 4.89]	3.86 [3.44, 4.29]	2.98 [2.76, 3.20]	1.00 (0.578)	4.01 (0.001)	4.53 (<0.001)
DOBJ.MEANPMI	1.54 [1.19, 1.90]	1.55 [1.39, 1.72]	1.09 [0.83, 1.35]	−0.07 (0.997)	2.33 (0.060)	3.44 (0.003)

modifiers with a higher PMI as compared to oral tasks. In the case of the L1 data, the only measure that was found to distinguish between the oral and written tasks was the number of adjectival modifier types. The fact that both L1 and learner written tasks had more adjectival modifier types as compared to oral tasks is likely related to the communicative function of the task. As shown in Example 1, the argumentative essays tended to be composed of long noun phrases, containing one or more adjectival modifiers. This is characteristic of texts with an "informational" communicative purpose (Biber, 2014). In contrast, both oral tasks are more verbal in nature. This is unsurprising given that both the oral narrative and the oral interview require participants to relate a sequence of events: either events in a story in the case of the narrative task (Example 2) or events occurring in their own lives in the case of the interviews (Example 3). As a result, the most important information in the oral tasks is conveyed through the use of verb phrases rather than complex noun phrases. This pattern is also in line with the register-variation literature showing that narrative functions such as these are associated with a more clausal style (Biber, 2014).

(1) *bien que c'est possible qu'une telle composition familiale puisse constituer une entrave pour l'enfant dans son développement il faut absolument note que ceci n'est pas forcément une conséquence du type de famille en soi même c'est à dire de l'orientation sexuelle des parents.*
'although it's possible that such a family composition can constitute an obstacle for the child in his development, it is necessary [to] note that this is not necessarily a consequence of the type of family itself, that is to say, the sexual orientation of the parents' (G126d).

(2) *et pendant ce temps là Jacques pensait à toutes les choses qu'ils faisaient ensemble. ils riraient ensembles. ils lisaient des histoires. et ils jouaient dans les arbres.*
'and during that time, Jacques thought of all of the things they did together. they laughed together. they read stories. and they played in the trees.' (B100c)

(3) *oui j'ai envoyé ma candidature à l'université. et j'ai choisi des cours que je veux faire. et. je vais continuer avec l'allemand la civilisation allemande. et j'espère aussi faire un cours qui s'agit de la langue alsacien.*
'yes, I sent in my application to the university. and I chose courses that I want to do. and. I am going to continue with German, German civilization. and I also hope to take a course about the Alsacian language.' (0118a).

In addition to the more varied use of adjectival modifier types, written learner productions also used adjectival modifiers with a higher PMI as compared to the oral productions. The most frequent adjectival modifiers in the written data include units such as *couple_NOM hétérosexuel_ADJ* ('heterosexual couple'; PMI = 6.57, n = 23) and *nourriture_NOM sain_ADJ* ("healthy food"; PMI = 4.49, n = 17), which are directly related to the essay topics. That may also explain why we found a slight topic effect for adjectival modifiers. Texts that contained vocabulary that was more similar to that of the gay marriage and adoption essay were found to contain slightly more adjectival modifier types (0.07) for every one standard deviation increase in COSINE.log. In other words, it appears that the argumentative essays promoted adjectival modifiers in general and certain topics required the use of specialized vocabulary, which promoted the use of a more diverse set of adjectival modifiers related to the topic. In the oral data, however, the most frequent adjectival modifiers tended to be more general and have a lower PMI: *temps_NOM même_ADJ* ('same time'; PMI = 3.12, n = 68), *chose_NOM autre_ADJ* ('other thing'; PMI = 2.31, n = 57). This contrasts somewhat with the native

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

data, which showed that the mean PMI of adjectival modifiers was relatively similar between the essay and the narrative and that the major difference was between both of those tasks and the interview. If access to appropriate, topic-specific adjectival modifier relations is somewhat more proceduralized for the native speakers, it may explain why the extra planning afforded by the written mode did not have a significant impact on the use of high PMI units by the native speakers but did have an impact for the learners in that it allowed learners to devote more attentional resources to selecting adjectival modifier relations (Ellis et al., 2019; Skehan, 1998). To that extent, these results are in line with previous research showing higher levels of both lexical diversity and sophistication (use of infrequent lexical items) in written tasks as compared to oral tasks in L2 French (Bulté & Housen, 2009; Granfeldt, 2007).

The reduced opportunity for online planning afforded by the oral tasks may also explain why those tasks exhibited less diversity of direct objects as compared to the written task. The pressure of online production seems to have favored the use and repetition of more frequent, "safe" direct objects containing high frequency verbs such as *faire* ('make/do') and *avoir* ('have'). In one interview for example, the direct object *faire_VER devoir_NOM* ('do homework') is repeated five times throughout. The phenomenon of learners sticking to a restricted set of highly frequent collocations, so-called lexico-grammatical teddy bears, has been previously attested in L2 written (Altenberg & Granger, 2001; Nesselhauf, 2005) and oral (Paquot et al., 2022) production. Because these highly frequent collocations often have a low PMI, this also has an effect on sophistication. While all three tasks exhibit direct objects with a negative PMI, indicating a lack of association (e.g., *changer_VER problème_NOM*; 'change problem'; PMI = −1.31), the large proportion of the direct objects with a low PMI (1–5) in the oral tasks seems to have counterbalanced the effect of the direct objects with negative PMI in those tasks. In other words, direct object sophistication is not significantly different between the oral and written tasks not because the direct objects used in the oral tasks are particularly sophisticated but, rather, because in pressured oral production, learners used (and repeated) a smaller set of safer, low PMI direct objects. Of course, recycling vocabulary can be a very useful communicative strategy, which may allow learners to increase fluency (see Dabrowska, 2014), but such an increase in fluency often comes at the expense of complexity (Skehan, 2009b), which is what the measures presented here aim to capture.

Putting everything together, the results show that phraseological complexity did indeed differ between oral and written tasks. In the case of the learners, this seems to be a result of both the functional characteristics of the task (e.g., the use of specific adjectival modifiers in the argumentative essays) as well as the opportunities for online planning afforded by the written modality. For the L1 participants, whose production is likely more proceduralized than the learners, phraseological complexity seems to be determined more so by the functional requirements of a task than modality per se.

RQ2. To what extent does the development of phraseological complexity differ between oral and written tasks completed by learners of French?

The raw phraseological complexity measures for all texts are plotted in Figure 1. The red lines in each graph show the estimated linear trends over time (with 95% confidence intervals) as predicted by the model and the blue dashed line shows the estimated marginal mean for the L1 group (95% confidence intervals indicated by gray shading). As shown by the figure, the two dimensions of phraseological complexity, namely diversity and sophistication showed different patterns of development. For diversity,

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

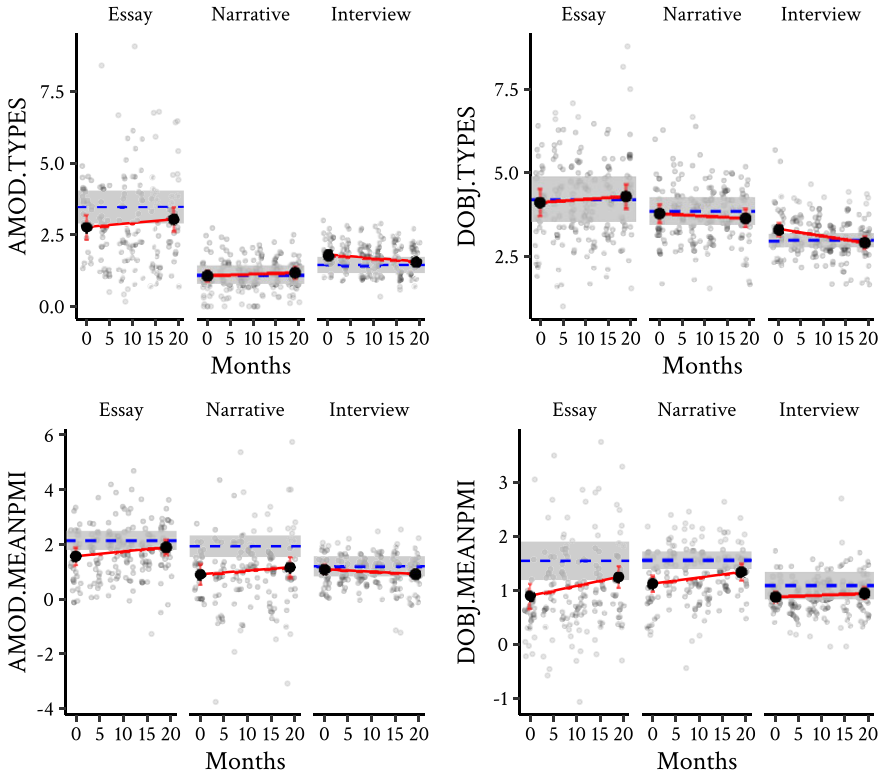

Figure 1. Predicted estimates and raw values of phraseological complexity measures over time (solid red line) compared to L1 benchmark (blue dashed line).

the only significant change was observed in the number of adjectival modifier types in the oral interviews, which decreased slightly over the study period. The fact that no significant increase was observed for phraseological diversity contrasts somewhat with the results of Qi and Ding (2011), the only longitudinal study to our knowledge that has measured the development of phraseological diversity, finding a significant increase over time in the diversity of manually identified "formulaic sequences" in oral monologues. However, that study was carried out over a longer period (four years) and involved a different type of phraseological unit so it is possible that the diversity of adjectival modifiers and direct objects simply develops too slowly to have been observed over the course of the 21-month period of the LANGSNAP project. That being said, given that the predicted phraseological diversity of the learners is generally within the same range as the L1 group, it may be that the learners had already mastered a level of complexity that was sufficient for these tasks. This may be why cross-sectional studies have also failed to show significant differences with respect to phraseological diversity between B2 and C2 level learners (Paquot, 2019) and between highly advanced learners and native speakers (Forsberg, 2010). In other words, this suggests that there is an upper limit on how phraseologically diverse a task needs to be, after which point an increase in proficiency is no longer associated with an increase in phraseological diversity.

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

In contrast to the diversity measures, which exhibited no clear growth, the sophistication of both direct objects and adjectival modifiers increased slightly over the study period. In the case of both essays and narratives, the PMI of units at the beginning of the study period was below the lower confidence interval for the L1 benchmark and moved closer toward the native benchmark over time. However, although the direction of the trend was positive for adjectival modifiers, it did not quite reach statistical significance, in contrast to the results of Edmonds and Gudmestad (2021), who found a significant increase in the PMI of adjective-noun collocations in the LANGSNAP argumentative essays between presojourn and the second postsojourn collection 19 months later. Although the Edmonds and Gudemestad study used the same data set, direct comparison of the results is made difficult due to several methodological differences including the use of a different extraction method, a different reference corpus and a different statistical modeling approach. That being said, one point of similarity between the two studies is that growth of phraseological sophistication is very gradual. Edmonds and Gudmestad (2021, p. 11) report a monthly increase of 0.04,[13] in the PMI of adjectival modifiers, compared to 0.01 (n.s.) in this study. Even when growth was found to be significant in the current study, in the case of direct objects, the change in phraseological complexity only amounted to a monthly increase in PMI of 0.01. This may also explain why previous research has often failed to find a significant effect of time on association strength. Yoon (2016), for example, reported no significant increase in the PMI of verb + noun collocations over the course of a semester in either narrative or argumentative written texts. Similarly, Kim, Crossley, and Kyle (2018) found no significant effect of time on the PMI of ngrams in learner interviews over the course of one year. These results suggest that phraseological complexity indeed develops very slowly, and that a large quantity of target language contact does not necessarily guarantee improvement in phraseological complexity (see Arvidsson, 2019).

Interestingly, although the difference in development between the three tasks was not significantly different for the PMI of direct objects, as can be seen in Figure 1, the estimated marginal trend for both the essay (0.02) and the narrative (0.01) are larger than that of the interview (0.003). In other words, direct object sophistication in both the essays and the narrative seems to have developed slightly faster than in the interviews (but not significantly so). One possible explanation may be that the interview topics simply did not elicit the same type of sophisticated vocabulary as the essay and narrative. Indeed, we did find that there was a small topic effect for the sophistication of direct objects so it could be the case that the open-ended nature of the interviews allowed learners to avoid topics for which they lacked vocabulary. Compared to the interview, the essay and the narrative tasks are both somewhat more constrained and this may have pushed the learners to produce more sophisticated phraseological units if they had acquired them in the intervening time between tasks. This was particularly evident in the case of one learner who struggled at presojourn to produce the word *conte* ("fairy tale") but a year later was able to produce the direct object relation *lire_VER conte_NOM* ("read fairy tale"; PMI = 2.89) when completing the same narrative task. Likewise, the same learner replaced the direct object *faire_VER exercise_NOM* ("do exercise"; PMI = 1.00) with the more sophisticated *chasser_VER papillon_NOM* ("'hunt butterfly"; PMI = 4.37), which describes the actions of the character in the story more specifically. Such gaps in vocabulary knowledge may not have been as

[13]Dividing the estimate from presojourn to postsojourn (0.7433) by the number of months between collection points (19).

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

evident in the interview task because there was less of a need to produce *specific* phraseological units and learners may have been able to recycle phraseological units used by the interlocutor or reuse units with which they are comfortable. Again, these may be very useful communication strategies in their own right but they have consequences for the ability to observe the development of phraseological complexity over time to the extent that they mask vocabulary gaps and artificially raise phraseological complexity at earlier time points (cf. Paquot et al., 2022).

To sum up, no significant increase was observed in terms of phraseological diversity over the 21-month period, which may have been because learners had already reached a native benchmark of phraseological diversity. Phraseological sophistication, however, did show a small amount of growth over the same period but this growth was only significant for direct objects. These results show that even in the context of intensive L1 input, phraseological complexity is slow to develop and that even when it does occur, it may not be equally evident in all task types or for all types of phraseological units.

Conclusion

While we first set out to compare phraseological complexity across modes, it quickly became apparent that a simple oral-written binary approach was too simplistic. While the tasks in the current study differ with respect to modality, they differ in other important ways as well. For one, the oral and written tasks involve different communicative functions (Biber, 2014), which we showed seem to have an influence on the type of phraseological units that are elicited. Tasks also differ with respect to performance conditions such as the amount of planning time and interactivity (Skehan, 2009b), both of which have been shown to have an influence on the complexity of learner productions (Ellis et al., 2019). The fact that phraseological complexity tended to be higher in learners' written tasks as compared to oral tasks could be due to a number of these task-related factors. In this study, the design of the corpus only allows us to speculate on the causes of the differences that we observed between the tasks. That being said, the comparison to native benchmark data did help to shed light on possible explanations.

If access to phraseological units is somewhat more proceduralized for native speakers as compared to the learners (De Bot, 1992), this may explain why writing seems to have showcased the phraseological complexity of learners more so than native speakers. Comparing the learner levels of phraseological diversity to the L1 benchmark also showed that the learners performed quite similarly to their L1 counterparts in some respects, which suggests that there is a limit to how phraseologically diverse a production needs to be to fulfill the requirements of a task. Similarly, although the rate of development was not significantly different between tasks, there was a general trend that suggested that development may not be equally evident in all tasks. If a task does not require sophisticated language, such language is not likely to be elicited from learners. To our knowledge, this is the first study to directly compare the effect of task type on phraseological complexity and the results speak to the importance of considering task variables when measuring phraseological complexity in L2 production. They also speak to the usefulness of including L1 benchmark data (cf. Housen & Kuiken, 2009).

It is important, however, to consider the limitations of this study when interpreting these results. For one, the assumption of linear growth over time may be criticized, given the dynamic nature of linguistic development (De Bot & Larsen-Freeman, 2011).

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

However, the main aim of this research was to determine the extent to which the longitudinal development of phraseological complexity differed across oral and written production tasks, so we feel that such an abstraction is merited. Although individual learners may have shown variability from any one time point to the next, the results here show that there was an overall increase in direct object PMI over time and this linear pattern did not show a large degree of variation between learners (evidenced by the small standard deviation in by-participant slopes for time in months). Moreover, approaches that fully capitalize on variability in L2 development, such as the dynamic systems approach (ibid.), require many more data collection points than the six that are available in the LANGSNAP data.

Another limitation is that to include both random slopes and random intercepts in the model, the individual prompts for the essays and narratives needed to be collapsed together. While this is not ideal, given that topic has been shown to have a strong effect on phraseological complexity (Paquot et al., 2021), we believe that the generalizability gained by including random slopes and random intercepts outweighs this downside. In an attempt to control for topic, we removed the phraseological units that were directly provided in the prompts or the picture stories and we included cosine similarity in the model. This method allowed us to quantify text differences in a more fine-grained way than is possible by using a simple categorical variable for the different prompts. That being said, it is still a rather crude measure of textual similarity. Future studies would do well to investigate more sophisticated methods of controlling for topic differences (e.g., topic modeling as proposed by Murakami et al., 2017).

The scope of this study is also limited in that we have chosen to focus on only two types of phraseological units (as realized in the form of adjectival modifiers and direct objects). Given that this was one of the first studies to investigate phraseological complexity in L2 French, we felt that looking at these units was a good starting point, especially in light of previous phraseological complexity research in L2 English (e.g., Paquot, 2018, 2019). Of course, it would be disingenuous to claim that these two units alone represent the full extent of the phraseological complexity apparent in a given text and more work is clearly needed to increase the coverage of phraseological complexity measures, especially considering that adjectival modifiers and direct objects tended to relatively infrequent in the learner texts overall. In oral narratives, for example, the median number of adjectival modifiers was only 1.18 per 100 words (IQR = 0.92). This raises the question of how many phraseological units are needed to get a reliable picture of a text's phraseological complexity. After all, if a text uses just one adjectival modifier with a high PMI, is that really enough to claim that it exhibits a high level of phraseological sophistication overall? Probably not. In a follow-up study to the present one, we have begun to address this very issue by broadening the scope of phraseological units under investigation beyond two-word collocations (Vandeweerd et al., under review).

Another challenge for phraseological complexity is how to operationalize the dimension of diversity. As discussed in the methodology section, measures that were originally designed for lexical items tend to be either strongly correlated with text-length overall (e.g., RTTR) or cannot be used because they require many more units than are attested in a typical text (e.g., MTLD). Using a moving-average method based on 100-word windows solved some of these problems but may have also introduced others (e.g., conflating low quantity and low diversity, biasing units at the beginning rather than the end of texts). Moving forward, it will be necessary to develop more innovative solutions to account for the unique challenges of phraseology rather than simply borrowing measures from other linguistic domains.

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

More broadly, because phraseological complexity is still a relatively new construct, there is also an urgent need to empirically demonstrate the validity of these measures more systematically. This could be done, for example, by grounding these measures in human judgements not just of proficiency but also of phraseological complexity (as suggested by Paquot, 2021). After all, just because a given complexity measure is correlated with proficiency or increases over time does not mean that it necessarily represents the construct it is intended to measure (see Pallotti, 2015, 2021).

Despite these limitations, the results here do speak to the usefulness of phraseological complexity measures as indices of development in L2 French. Importantly, they also show that various factors besides development appear to influence the amount of phraseological complexity observed in a given text. These factors are similar to those discussed by Skehan (2009a) in relation to *lexical complexity*, namely: performance conditions (e.g., modality), style, and other task influences (e.g., interactivity). At this point, we have only just begun to scratch the surface in our understanding of how such factors contribute to measures of phraseological complexity and this will be an important area to explore further in the future.

Acknowledgments. We would like to thank Héloïse Copin for her invaluable help with the manual annotation. We are also grateful to the LANGSNAP team at the University of Southampton who have kindly made their data publicly available for research purposes. Funding for this project was provided by the Fonds de la Recherche Scientifique—FNRS (Grant n° T.0086.18). The present research also benefited from computational resources made available on the Tier-1 supercomputer of the Fédération Wallonie-Bruxelles, infrastructure funded by the Walloon Region under the grant agreement n°1117545.

Data Availability Statement. The experiment in this article earned an Open Materials badge for transparent practices. The materials are available at: https://osf.io/bkfru/

References

Abeillé, A., & Barrier, N. (2004). Enriching a French treebank. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.), *Proceedings of the fourth international conference on language resources and evaluation (LREC'04)* (pp. 2233–2236). European Language Resources Association.

Abeillé, A., & Clément, L. (2003). *Annotation morpho-syntaxique.* Université Paris 7. http://www.llf.cnrs.fr/sites/sandbox.linguist.univ-paris-diderot.fr/files/statiques/french_treebank/guide-morpho-synt.02.pdf

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

Arvidsson, K. (2019). Quantity of target language contact in study abroad and knowledge of multiword expressions: A usage-based approach to L2 development. *Study Abroad Research in Second Language Acquisition and International Education*, *4*, 145–167.

Bestgen, Y., & Granger, S. (2018). Tracking L2 writers' phraseological development using collgrams: Evidence from a longitudinal EFL corpus. In S. Hoffmann, A. Sand, S. Arndt-Lappe, & L. M. Dillmann (Eds.), *Corpora and lexis* (pp. 277–301). Brill Rodopi.

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*, *14*, 7–34.

Biber, D. (2019). Text-linguistic approaches to register variation. *Register Studies*, *1*, 42–75.

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

Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the Toefl iBT test: A lexico-grammatical analysis*. Educational Testing Service.

Blanche-Benveniste, C., & Adam, J.-P. (1999). La conjugaison des verbes: Virtuelle, attestée, defective. *Recherches sur le français parlé*, *15*, 87–112.

Bulté, B., & Housen, A. (2009). *The development of lexical proficiency in L2 speaking and writing tasks by Dutch-speaking learners of French in Brussels.* Paper presented at the Task Based Language Teaching Conference, Lancaster, UK.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, I. Vedder, & F. Kuiken (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). John Benjamins.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42–65.

Candito, M., Nivre, J., Denis, P., & Anguiano, E. H. (2010). Benchmarking of statistical dependency parsers for French. In C.-R. Huang & D. Jurafsky (Eds.), *COLING 2010: Poster volume* (pp. 108–116). Coling 2010 Organizing Committee.

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

Covington, M. A., & McFall, J. D. (2010). Cutting the gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*, 94–100.

Dabrowska, E. (2014). Recycling utterances: A speaker's guide to sentence processing. *Cognitive Linguistics*, *25*, 617–653.

De Bot, K. (1992). A bilingual production model: Levelt's "speaking" model adapted. *Applied Linguistics*, *13*, 384–404.

De Bot, K., & Larsen-Freeman, D. (2011). Researching second language development from a dynamic systems theory perspective. In M. Verspoor, K. de Bot, & W. Lowie (Eds.), *A dynamic approach to second language development: Methods and techniques* (pp. 5–24). John Benjamins.

DeKeyser, R. (2007). Study abroad as foreign language practice. In R. DeKeyser (Ed.), *Practice in a second language* (pp. 208–226). Cambridge University Press.

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.

Edmonds, A., & Gudmestad, A. (2021). Collocational development during a stay abroad. *Languages*, *6*, 1–17.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics and TESOL. *TESOL Quarterly*, *5*, 375–396.

Ellis, R., Skehan, P., Li, S., Shintani, N., & Lambert, C. (2019). Psycholinguistic perspectives. In R. Ellis, P. Skehan, S. Li, N. Shintani, & C. Lambert (Eds.), *Task-based language teaching: Theory and practice* (pp. 64–102). Cambridge University Press.

Ellis, R., & Yuan, F. (2005). The effects of careful within-task planning on oral and written task performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 167–192). John Benjamins.

Forsberg, F. (2010). Using conventional sequences in L2 French. *International Review of Applied Linguistics in Language Teaching*, *48*, 25–51.

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

Granfeldt, J. (2007). Speaking and writing in French L2: Exploring effects on fluency, complexity and accuracy. In S. van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning & teaching* (pp. 87–98). Koninklijk Vlaamse Academie van Belgie voor Wetenschappen en Kunsten.

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., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 27–50). John Benjamins.

Gries, S. Th. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 3–25). John Benjamins.

Guiraud, P. (1954). *Les charactères statistiques du vocabulaire*. Presses Universitaires de France.

Holliday, A. (2006). Native-speakerism. *ELT Journal*, *60*, 385–387.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*, 461–473.

Housen, A., Kuiken, F., & Vedder, I. (2012). *Dimensions of L2 performance and proficiency: Complexity, Accuracy and Fluency in SLA*. John Benjamins.

Kim, M., Crossley, S. A., & 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.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, *1*, 60–69.

Kormos, J. (2014). Differences across modalities of performance: An investigation of linguistic and discourse complexity in narrative tasks. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 193–216). John Benjamins.

Kuiken, F., & Vedder, I. (2011). Task complexity and linguistic performance in L2 writing and speaking. In *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 91–104). John Benjamins.

Kyle, K., & Eguchi, M. (2021). Automatically assessing lexical sophistication using word, bigram, and dependency indices. In S. Granger (Ed.), *Perspectives on the L2 phrasicon: The view from learner corpora* (pp. 126–151). Multilingual Matters.

Larsen-Freeman, D., & Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning*, *27*, 123–134.

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

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk*. 3rd Edition. Lawrence Erlbaum Associates.

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.

Mitchell, R., Tracy-Ventura, N., & McManus, K. (2017). *Anglophone students abroad*. Routledge.

Murakami, A., Thompson, P., Hunston, S., & Vajn, D. (2017). "What is this corpus about?" Using topic modelling to explore a specialised corpus. *Corpora*, *12*, 243–277.

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

Nivre, J., Hall, J., & Nilsson, J. (2006). MaltParser: A data-driven parser-generator for dependency parsing. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Eds.), *Proceedings of the fifth international conference on language resources and evaluation (LREC'05)* (pp. 2216–2219). European Language Resources Association.

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

Ortega, L. (2019). SLA and the study of equitable multilingualism. *Modern Language Journal*, *103*, 23–38.

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

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*, 117–134.

Pallotti, G. (2021). Measuring complexity, accuracy, and fluency (CAF). In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 201–210). Routledge.

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

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*, 121–145.

Paquot, M. (2021). *Measures of phraseological complexity: Reliability and validity.* Paper presented at the World Congress of Applied Linguistics, Groningen, The Netherlands.

Paquot, M., Gablasova, D., Brezina, V., & Naets, H. (2022). Phraseological complexity in EFL learners' spoken production across proficiency levels. In A. Lénko-Szymańska & S. Götz (Eds.), *Complexity, accuracy and fluency in learner corpus research* (pp. 115–136). John Benjamins.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130–149.

Paquot, M., Naets, H., & Gries, S. Th. (2021). Using syntactic co-occurrences to trace phraseological complexity development in learner writing: Verb + object structures in LONGDALE. In B. Le Bruyn & M. Paquot (Eds.), *Learner corpus research meets second language acquisition* (pp. 122–147). Cambridge University Press.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2020). *nlme: Linear and Nonlinear Mixed Effects Models.* (Version 3.1-149) [Computer software]. https://cran.r-project.org/package=nlme

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, *100*, 538–553.

https://doi.org/10.1017/S0272263122000389 Published online by Cambridge University Press

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research. *Language Learning*, *65*, 37–75.

Qi, Y., & Ding, Y. (2011). Use of formulaic sequences in monologues of Chinese EFL learners. *System*, *39*, 164–174.

Ravid, D., & Tolchinsky, L. (2002). Developing linguistic literacy: A comprehensive model. *Journal of Child Language*, *29*, 417–447.

R Core Team. (2021). *R: A language and environment for statistical computing.* (Version 4.1.0) [Computer software]. https://www.r-project.org/

Rubin, R., Housen, A., & Paquot, M. (2021). Phraseological complexity as an index of L2 Dutch writing proficiency: A partial replication study. In S. Granger (Ed.), *Perspectives on the second language phrasicon: The view from learner corpora* (pp. 101–125). Multilingual Matters.

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, & A. Witt (Eds.), *Proceedings of challenges in the management of large corpora 3 (CMLC-3)* (pp. 28–34). Institut für Deutsche Sprache.

Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 486–493). European Language Resources Association.

Schmid, H. (1994). *TreeTagger: A part-of-speech tagger for many languages.* [Computer software]. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

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

Siyanova-Chanturia, A. (2015). Collocation in beginner learner writing: A longitudinal study. *System*, *53*, 148–160.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Skehan, P. (2009a). Lexical performance by native and non-native speakers on language-learning tasks. In B. H. Richards, H. Daller, D. 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. 107–124). Palgrave Macmillan.

Skehan, P. (2009b). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*, 510–532.

Skehan, P. (2014). Limited attentional capacity, second language performance, and task-based pedagogy. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 211–260). John Benjamins.

Skehan, P., & Foster, P. (2008). Complexity, accuracy, fluency and lexis in task-based performance: A synthesis of the Ealing research. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. (pp. 199–220). University of Brussels Press.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44*, 307–336.

Tracy-Ventura, N., McManus, K., Norris, J. M., & Ortega, L. (2014). "Repeat as much as you can": Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 143–163). Multilingual Matters.

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). John Benjamins.

Vandeweerd, N., Housen, A., & Paquot, M. (2021). Applying phraseological complexity measures to L2 French: A partial replication study. *International Journal of Learner Corpus Research*, *7*, 197–229.

Vandeweerd, N., Housen, A., & Paquot, M. (under review). Proficiency at the lexis-grammar interface: Comparing oral versus written French exam tasks.

Wang, J., & Dong, Y. (2020). Measurement of text similarity: A survey. *Information (Switzerland)*, *11*, 1–17.

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

Zhang, X., & Li, W. (2021). Effects of n-grams on the rated L2 writing quality of expository essays: A conceptual replication and extension. *System*, *97*, 102437.

Appendix

Definitions of phraseological units

- Adjectival modifiers (AMOD) were defined as adjectives that modify a noun. This also includes superlatives (e.g., la *drogue* la plus *forte*). Following Abeillé and Clément (2003), past participles in the attribute position are considered adjectives (e.g., *enfants adoptés*) except if they are followed by an agentive complement (e.g., *les enfants adoptés par Jean et Luc*). Present participles are considered adjectives if they agree with the noun they modify and they do not have a direct object (e.g., *les erreurs existants*). Quantifiers (e.g., *certaines*) and ordinal numbers (e.g., *première*) are not considered adjectives, nor are nouns that are modified by another (proper) noun (e.g., *étudiants Erasmus*). Compound words (e.g., *fast food*) are not considered adjectival modifiers.
- Direct objects (DOBJ) were defined as nouns that are the object of verbs. This does not include nouns modified by a relative clause (e.g., les *impôts* qu'on *paie*) or by a passive clause (les *distributeurs* ont été *supprimés*) but it does include objects of nonfinite verbs (e.g., *concernant* le *mariage*). The object of *il y a* is considered a direct object relation (e.g., il y *a* une bonne *idée*).

Cite this article: Vandeweerd, N., Housen, A. and Paquot, M. (2023). Comparing the longitudinal development of phraseological complexity across oral and written tasks. *Studies in Second Language Acquisition,* 45: 787–811. https://doi.org/10.1017/S0272263122000389