Applying Phraseological Complexity Measures to L2 French:
A Partial Replication Study

Nathan Vandeweerd[1,2], Alex Housen[2] and Magali Paquot[1]
[1]Université catholique de Louvain / [2]Vrije Universiteit Brussel

This study partially replicates Paquot's (2018, 2019) study of phraseological complexity in L2 English by investigating how phraseological complexity compares across proficiency levels as well as how phraseological complexity measures relate to lexical, syntactic and morphological complexity measures in a corpus of L2 French argumentative essays. Phraseological complexity is operationalized as the diversity (root type-token ratio; RTTR) and sophistication (pointwise mutual information; PMI) of three types of grammatical dependencies: adjectival modifiers, adverbial modifiers and direct objects. Results reveal a significant increase in the mean PMI of direct objects and the RTTR of adjectival modifiers across proficiency levels. In addition to phraseological sophistication, important predictors of proficiency include measures of lexical diversity, lexical sophistication, syntactic (phrasal) complexity and morphological complexity. The results provide cross-linguistic validation for the results of Paquot (2018, 2019) and further highlight the importance of including phraseological measures in the current repertoire of L2 complexity measures.

Keywords: L2 French; replication, phraseology; collocations; complexity; CEFR

1. **Introduction**

It is well-recognized that complexity plays an important role in the development of L2 proficiency along with accuracy and fluency (Skehan, 2009). To date, however, most research on complexity has focused on isolated linguistic domains, with a particular focus on lexical and syntactic complexity (Bulté & Housen, 2012) and scant research focusing on complexity at the lexis-grammar interface, despite theoretical motivations for considering lexis and grammar as part of the same continuum (see e.g., Goldberg, 2006; Hunston & Francis, 2000). Recently, a new line of research has started to examine complexity at this interface by investigating the development of phraseological complexity, that is to say, the diversity and sophistication of phraseological units (Paquot, 2019).

This line of research is inspired by L2 phraseology research, which has shown that as learners become more advanced, they tend to use more infrequent and highly exclusive word combinations. This can be measured using pointwise mutual information (PMI), which quantifies the probability of co-occurrence, given the respective frequencies of two individual words (see Church & Hanks, 1990). Durrant and Schmitt (2009) for example, found that texts written by native writers had more collocations with a high PMI , suggesting that native writers were more sensitive to rare but highly collocated word pairs. Similarly, Granger and Bestgen (2014) found that texts written by advanced learners had a significantly higher proportion of collocations with a high PMI score. In particular, the sophistication of adjective-noun collocations was found to be the best discriminator between learners at the B and C levels of the Common European Framework of Reference (Council of Europe, 2001). These findings were also supported by a follow-up study which looked at longitudinal phraseological development and found that there was an increase in the proportion of high PMI collocations over time (Bestgen & Granger, 2018). Garner, Crossley and Kyle (2018) also recently found that association strength was a significant predictor of the proficiency level of the learner texts in their corpus.

According to Paquot (2019), measures of association strength represent the depth or sophistication of knowledge that a learner has about word combinations. In addition to depth, Paquot argued that this knowledge should also be measured in terms of breadth, following common operationalizations of complexity in other linguistic domains (Bulté & Housen, 2012). In the domain of lexical complexity, the dimension of breadth is usually operationalized in terms of the diversity of different words used by a learner. Following this logic, Paquot (2019: 124) defined phraseological complexity as: "the range of phraseological units that surface in language production and the degree of sophistication of such phraseological units". The phraseological units in question were dependency relations: binary relationships between a head and its dependent which are obtained using a dependency parser which establishes these binary pairs on the basis of statistical extrapolation from a set of manually annotated syntactic trees. Paquot (2018, 2019) explored three types of dependency relations: adjectival modifiers (AMOD; e.g. *black* + *hair*), adverbial modifiers (ADVMOD; e.g. *very* + *black*) and direct objects (DOBJ; e.g. *win* + *lottery*). Diversity was operationalized as the root type-token ratio of the dependency relations (Paquot, 2019) and sophistication as the mean PMI score (Paquot, 2018, 2019) and the proportion of the dependency units in four collocational bands (Paquot, 2018). Using a corpus of linguistics essays written by L1 French EFL learners, Paquot (2019) found that the mean PMI of adjectival modifiers could better predict the proficiency level of the texts than traditional lexical or syntactic measures. A follow-up study also showed that the mean PMI of direct objects and adjectival modifiers in particular explained 25% of the variance in holistic ratings of the essays (Paquot, 2018).[1] The diversity of the phraseological units was

[1] The final model did not include the mean PMI of adverbial modifiers.

not found to be predictive of the Common European Framework of Reference (CEFR) level (Council of Europe, 2001) in either study.

Until now, phraseological complexity research has focused on L2 English (see Rubin, Housen and Paquot, in press, for L2 Dutch) but there is evidence to suggest that phraseological complexity may be slower to develop in more synthetic languages, such as French. Stengers, Boers, Housen et al. (2011) compared the effect of formulaic language on holistic assessments of oral proficiency in L2 English and L2 Spanish and found that the correlation of what they called formulaic language use to proficiency ratings was weaker for the Spanish learners than for the English learners. The authors suggest that the greater inflectional demands of learning L2 Spanish outweighed the contribution of formulaic language. Compared to English, where content word forms have relatively few morphological variants, in a highly inflected language, learners need to acquire many more forms of the same collocation (for example, multiple verbal and nominal inflections of verb-noun collocations). There is therefore reason to believe that phraseological complexity may develop more slowly in synthetic languages and exhibit trade-offs with morphological complexity. Thus the main aim of this paper is to determine how phraseological complexity develops in a more synthetic language by partially replicating the results of Paquot (2018, 2019) on L2 English on a comparable sample of L2 learners of French.

2. Complexity research in L2 French

Complexity in traditional linguistic domains has shown to be associated with increased proficiency in L2 French. Lexical diversity, operationalized as the root type-token ratio, has been shown to distinguish between proficiency levels (De Clercq, 2015) as has lexical sophistication in terms of the proportion of low frequency words (De Clercq, 2015; Ovtcharov, Cobb, & Halter, 2006). Similar relationships with proficiency have been found for syntactic complexity in terms of length of syntactic units (De Clercq & Housen, 2017; Gyllstad, Granfeldt, Bernardini, & Källkvist, 2014) and in terms of amount of juxtaposition, coordination

(Welcomme, 2013) and subordination (De Clercq & Housen, 2017; Gyllstad et al., 2014; Welcomme, 2013). At the phrasal level, De Clercq and Housen (2017) found no significant difference in noun phrase length between four groups of high school learners of French but there was a significant difference between the learner groups and a native-speaker control, suggesting that noun phrase complexity may play a role at more advanced levels of L2 French. Only one study to our knowledge has explicitly compared the development of complexity in different linguistic domains in L2 French. Over a series of studies, De Clercq (2016) studied the development of lexical, syntactic and morphological complexity in high school learners of French and English. The learners were grouped into four proficiency groups based on age and accuracy measures. The results showed that all three types of complexity increased in parallel with proficiency in both L2 French and L2 English and no trade-offs were observed between complexity domains at the group-level (but see Bulté, 2013; Verspoor, Schmid, & Xu, 2012 for trade-offs at the individual level in L2 English). The timing of complexity development did, however, differ. Whereas lexical diversity increased continuously across all four proficiency levels, the group-differences for lexical sophistication and morphological complexity were largest between the first two proficiency levels and the group-differences for syntactic complexity (at the level of the AS-unit) were largest between the middle two proficiency levels. However, it is important to keep in mind that most studies of complexity in L2 French have been primarily based on oral productions (cf. Gyllstad et al., 2014) so it is unclear whether similar developmental trends would be found in the written mode. Written French differs particularly with respect to morphology as there are many morphological inflections which have identical phonological realizations but are orthographically distinct (e.g. *regarde*; see-3SG.PRS.IND versus *regardent*; see-3PL.PRES.IND) (see Blanche-Benveniste & Adam, 1999). Such cross-modal differences must therefore be taken into account when considering complexity in L2 written French, which is the focus of the current study.

Several studies have also shown links between phraseology and proficiency in L2 French. Forsberg and Bartning (2010), for example, showed that the number of lexical formulaic sequences (e.g. *je vous en prie;* 'you're welcome') significantly increased between the A, B and C levels of the CEFR. Forsberg Lundell et al. (2018) also showed that productive knowledge of verb-object collocations was significantly higher at C1 as compared to B2 levels on a cloze test. De Clercq (2015) calculated the ratio of all words in a text to the number of words in verb-(preposition)-noun collocations with a PMI score above 3 and found a linear increase with proficiency, showing that more advanced learners use more strongly associated verb-noun combinations. In their study of long-residency learners, Erman, Denke, Fant and Forsberg Lundell (2015) found that even very advanced L2 French speakers, those who had lived on average ten years in France, showed significantly less diversity in their use of formulaic expressions when compared to native speakers, in contrast to a comparable group of English learners who were indistinguishable from native speakers in this regard, suggesting that phraseological diversity may be slower to develop in L2 French compared to L2 English. These studies show that as learners of French become more advanced, they use a higher quantity of highly-collocated word combinations but that even advanced learners tend use a more limited range of word combinations than native speakers. Exactly how phraseological complexity develops across the two dimensions of diversity and sophistication is still unclear as is the relationship between phraseological measures and so-called traditional measures of complexity. Our main research questions are therefore:

RQ1. How does phraseological complexity compare in written L2 French at different proficiency levels?

RQ2. To what extent does phraseological complexity relate to lexical, syntactic and morphological complexity in L2 written French?

3. Data and Method:

According to Porte (2012: 3), replication research "attempts to discover whether the same findings are obtained by another researcher in another context". In this case, replicating the methodology of Paquot (2018, 2019) for L2 English in the context of L2 French can provide insight into the generalizability of phraseological complexity measures to more synthetic languages. As a replication study, the main research questions and methodology have been borrowed from Paquot (2018, 2019) and every attempt was made to maintain consistency with the methods used, changing only the population: substituting learners of French for learners of English. That being said, changing the learner population presents several challenges. For example, using a different learner population also requires the use of different reference corpora and the availability of automatic linguistic tools is not the same across languages. We have therefore substituted comparable measures (based on comparable reference corpora) or developed our own tools where necessary to fill the gap, which means that this study is considered a partial or approximate replication (see Porte, 2012: 8). All differences between the methods used by Paquot (2018, 2019) and the current study are highlighted in the following sections.

3.1. Learner Data

As there is currently no corpus of L2 French research papers, we have instead used a corpus of argumentative essays from university students : the *Leerdercorpus Frans* (Demol & Hadermann, 2008; Vanderbauwhede, 2012). Complexity features can vary between tasks and registers of academic writing (e.g. Staples, Egbert, Biber, & Gray, 2016). As such, the comparison of the results to those of Paquot (2018, 2019) will need to take into account the difference in register and text length in addition to target language.

The *Leerdercorpus Frans* contains texts written by L1 Dutch learners of French, enrolled in their first or second year of the program “Language and Literature” at universities in Dutch-speaking Belgium. In addition to studying French, each student also specialized in one additional language. At the time of data collection, the learners had been studying French for approximately eight years. The corpus contains 253 argumentative essays written about seven different topics (see Table 1).[2] This study only included texts longer than 100 words as measures of lexical diversity are known to be unreliable for texts shorter than this. Some students contributed multiple texts to the corpus so we used a random sample to select only one text per writer. As a result, the corpus used in the analysis below is made up of 169 texts (84 888 word tokens). On average, the texts are about 502.3 words long (SD = 157.03), which is shorter than the texts in Paquot (2018, 2019) which were 3000-3500 words on average.

Table 1. Argumentative writing topics

Writing topics	
1.	La loi sur l’euthanasie doit-elle également s’appliquer aux mineurs (d’âge)? *Should the euthanasia law also apply to minors ?*
2.	Faut-il maintenir les centrales nucléaires au-delà de 2015? *Should we maintain nuclear power plants beyond 2015?*
3.	Une nouvelle réforme de l’État est-elle une priorité? *Is reforming the state a priority?*
4.	Que faire des jeunes délinquants ? *What should be done with delinquent youths?*
5.	La liberté est le droit de faire tout ce que les lois permettent (Montesquieu). *Liberty is the right to do anything the laws permit (Montesquieu).*
6.	Une langue sans professeurs, c’est comme une justice sans juges, un contrat sans notaire (Claudel). *Language without teachers is like the law without judges, a contract without lawyers (Claudel).*
7.	Les pouvoirs publics doivent-ils engager un pourcentage minimum d’allochtones? *Should the government be forced to hire a minimum percentage of immigrants?*

Each text was rated for proficiency by two trained raters with experience evaluating French proficiency exams according to the CEFR. The texts were assigned a CEFR level for each of the following criteria: grammatical accuracy, vocabulary control, vocabulary range, orthographic control, cohesion and coherence as well as a global CEFR level but only the

[2] The essays were assigned as part of regular coursework at two different universities. For three assignments, the topics were fixed (3, 5, 6) but for the other assignments, the students were free to select the topic of their choice from a subset (1, 2, 4, 7). With the exception of topic 4, all essays were written in the second year of the bachelors program but because the texts were subsequently evaluated for proficiency, the topics themselves are not tied to any specific proficiency level.

global proficiency score was used for the subsequent analysis following Paquot (2019). The raters reached a high level of agreement on this score (κ= 0.837, p < 0.001). Texts which did not receive the same global level by both raters were resubmitted to the raters to be reassessed and were eliminated from analysis if no agreement could be reached after the second round of assessment. Table 2 lists the number of tokens at each level within the final corpus.

Table 2. Assessed proficiency levels of the learner corpus

	Number of texts	Tokens	Mean length (tokens)	SD length (tokens)
B2	26	13348	513.38	155.3
C1	106	51981	490.39	164.44
C2	37	19559	528.62	135.06
Corpus	169	84888	502.3	157.03

N.B. A Kruskal-Wallis test revealed that the differences in number of tokens between the three proficiency levels was not significant (H(2) = 1.5758, p = 0.4548)

3.2. Complexity Measures

3.2.1. Phraseological Complexity

As in Paquot (2018, 2019), phraseological complexity was operationalized as the diversity and sophistication of three types of dependency relations: adjectival modifiers (AMOD; adjective + noun), adverbial modifiers (ADVMOD; adverb + verb) and direct objects (DOBJ; verb + direct object). In order to extract the dependencies, the learner texts were part-of-speech (POS) tagged with MElt POS Tagger (Denis & Sagot, 2012) and dependency parsed using Malt Parser (Candito, Nivre, Denis, & Anguiano, 2010). As these natural language processing (NLP) tools were originally developed for L1 French, we first established their reliability on learner language. To this end, 100 sentences were randomly extracted from the learner corpus and were manually annotated by two annotators for the three types of dependencies described above. The two annotators reached a high level of agreement (κ= 0.88, p < 0.001). All cases of disagreement were discussed and resolved in order to come up with a gold standard which could be compared to the output of the automatic parser. When compared to the manual annotation, we obtained an F1 score of 0.80 or greater for the automated annotation of each dependency relation of interest (see Table 3).

Table 3. Precision and recall scores for automatically identified dependencies

	Precision	Recall	F1
AMOD	0.90	0.73	0.81
ADVMOD_V	0.73	0.88	0.80
DOBJ	0.90	0.75	0.82

Once extracted, dependency tags were then reformatted using in-house Python scripts (Van Rossum & Drake, 2009). Phraseological diversity was calculated using the *koRpus* package in R (Michalke, 2019; R Core Team, 2019) as well as in-house R scripts for each learner text as the root type-token ratio of each dependency type. In order to reduce the possibility that dependency pairs containing spelling mistakes or that words which were incorrectly tagged by the POS tagger would be counted as unique, the calculation only included dependency pairs which occurred more than five times in the reference corpus, which was the *FRCOW16 corpus* (Schäfer, 2015; Schäfer & Bildhauer, 2012), a web-scraped corpus which contains approximately 10 billion words from French-language internet domains and which provides syntactic annotations with dependency parses obtained using the same processing chain which we used for the learner corpus. The same reference corpus was used to calculate a PMI score for each dependency pair found in the learner corpus based on the frequency of each word separately and their combined frequency (for details see Paquot, 2019). Again, only dependency pairs which occurred more than five times in the reference corpus were included in the calculation. Dependency pairs which occurred in the writing prompts were also excluded from the calculations. A mean PMI score for each dependency type was then calculated for each learner text. Following Paquot (2018), we also calculated the proportion of dependency pairs in each text which fell into four collocation bands: non-collocating (MI < 3), low (MI $3 \leq 5$), medium (MI $5 \leq 7$), and high (MI > 7). Table 4 lists the phraseological measures.

Table 4. Measures of Phraseological Complexity

Measure	Formula
Diversity	
Root TTR for AMOD dependencies	Tamod/√Namod
Root TTR for ADVMOD_V dependencies	Tadvmod√Nadvmod
Root TTR for DOBJ dependencies	Tdobj/√Ndobj
Sophistication: Mean PMI	
Mean PMI for AMOD dependencies	Σ MIamod / Namod
Mean PMI for ADVMOD_V dependencies	Σ MIadvmod/ Nadvmod
Mean PMI for DOBJ dependencies	Σ MIdobj / Ndobj
Sophistication: Collocation Bands	
Proportion of non-coll AMOD dependencies	Σ (MIamod <3) / Namod
Proportion of low-coll AMOD dependencies	Σ (MIamod 3 ≤ 5) / Namod
Proportion of med-coll AMOD dependencies	Σ (MIamod 5 ≤ 7) / Namod
Proportion of high-coll AMOD dependencies	Σ (MIamod > 7) / Namod
Proportion of non-coll ADVMOD_V dependencies	Σ (MIadvmod <3) / N advmod
Proportion of low-coll ADVMOD_V dependencies	Σ (MIadvmod 3 ≤ 5) / Nadvmod
Proportion of med-coll ADVMOD_V dependencies	Σ (MIadvmod 5 ≤ 7) / Nadvmod
Proportion of high-coll ADVMOD_V dependencies	Σ (MIadvmod > 7) / Nadvmod
Proportion of non-coll DOBJ dependencies	Σ (MIdobj <3) / Ndobj
Proportion of low-coll DOBJ dependencies	Σ (MIdobj 3 ≤ 5) / Ndobj
Proportion of med-coll DOBJ dependencies	Σ (MIdobj 5 ≤ 7) / Ndobj
Proportion of high-coll DOBJ dependencies	Σ (MIdobj > 7) / Ndobj

3.2.2. Lexical Complexity

In order to maximize comparability with Paquot (2018, 2019) as well as to avoid having multiple measures for the same construct (Ortega, 2012), lexical diversity was operationalized as Guiraud's (1954) root type-token ratio (RTTR). Following Paquot (2018, 2019), we calculated the diversity of all lexical words together (nouns, verbs, adjectives and adverbs), modifiers (adjectives and adverbs) as well as nouns, verbs, adjectives and adverbs separately. As suggested by Treffers-Daller (2013), all measures were calculated on lemmas, given the highly inflected nature of French. Also following Paquot (2018, 2019), we calculated the diversity of each word class as a proportion of all lexical lemmas and as a RTTR of the specific word class. These measures were calculated using the *koRpus* package (Michalke, 2019) for R (R Core Team, 2019) as well as in-house R scripts.

Further following Paquot (2018, 2019), we operationalized lexical sophistication as the proportion of words absent from a list of the 2000 most frequent words. As there is no comparable corpus of French with the same size and range of text types as the British National

Corpus used by Paquot, our measures are instead based on a French frequency dictionary (FFD) (Lonsdale & Le Bras, 2009), a 5000-word vocabulary list for French compiled from a 23-million word spoken and written corpus of French containing subtitles, literature, parliamentary proceedings, telephone conversations, theatre scripts, newspapers, popular science articles and technical reports. Thus far, no study has made use of this frequency list to measure lexical sophistication in learner productions but two separate vocabulary tests developed using the frequency data from this list have shown to discriminate between proficiency levels (Batista & Horst, 2016; Peters, Velghe, & Van Rompaey, 2019). As in Paquot (2018, 2019), the FFD-based measures were corrected for text-length by applying Giraud's correction. Because this list has not yet been used for measuring lexical sophistication in L2 French, we decided to complement this measure with a measure targeting advanced vocabulary as well. Studies in L2 English have made use of the Academic Word List (Coxhead, 2000) to measure advanced vocabulary. Although no such list currently exists for French, Tutin and Grossman (2014) recently developed the Lexique Transdisciplinaire (LT), a list of "transdisciplinary words" from a corpus of scientific articles. Each word on the list appears at least 15 times in the subdisciplines of linguistics, economics and medicine. This may be a possible equivalent to the English AWL but no study thus far has used this list to measure lexical sophistication in learner texts. This measure should therefore also be seen as an exploratory measure in the current study. To summarize, lexical sophistication was operationalized as the proportion of lexical lemmas (nouns, verbs, adjectives and adverbs) not appearing in the top 2000 most frequent words of the FFD as well as the proportion of verb, noun and adjective lemmas appearing on the LT list. Content words which occurred in the prompts were excluded from calculations of sophistication as well as those tagged by MElt tagger as proper nouns or unknown words. Table 5 lists the lexical complexity measures.

Table 5. Lexical complexity measures (T = type, N = token). Sophisticated lemmas defined as lemmas not appearing in the top 2000 most frequent lemmas of the French Frequency Dictionary (FFD; Lonsdale & Le Bras, 2009) and those appearing on the Lexique Transdisciplinaire list (LT; Tutin & Grossman, 2014)

Measure	Formula
Lexical Diversity	
Root TTR	T/√N
Lexical word variation	Tlex/Nlex
Noun variation	Tnoun/Nlex
Verb variation	Tverb/Nlex
Adjective variation	Tadj/Nlex
Adverb variation	Tadv/Nlex
Modifier variation	Tadv+adj/Nlex
Root TTR of nouns	Tnouns/√Nnouns
Root TTR of verbs	Tverbs/√Nverbs
Root TTR of adjectives	Tadj/√Nadj
Root TTR of adverbs	Tadv/√Nadv
Lexical Sophistication	
Proportion of FFD off-list lexical lemmas	$T_{FFD\text{-}offlist}lex/\sqrt{Nlex}$
Proportion of FFD off-list verb lemmas	$T_{FFD\text{-}offlist}verb/\sqrt{Nverb}$
Proportion of FFD off-list noun lemmas	$T_{FFD\text{-}offlist}noun/\sqrt{Nnoun}$
Proportion of FFD off-list adjective lemmas	$T_{FFD\text{-}offlist}adj/\sqrt{Nadj}$
Proportion of FFD off-list adverb lemmas	$T_{FFD\text{-}offlist}adv/\sqrt{Nadv}$
Proportion of LT on-list verb lemmas	$N_{LT\text{-}onlist}verb/Nverb$
Proportion of LT on-list noun lemmas	$N_{LT\text{-}onlist}noun/Nnoun$
Proportion of LT on-list adjective lemmas	$N_{LT\text{-}onlist}adj/Nadj$

3.2.3. Syntactic Complexity

For each syntactic level (sentence, t-unit, clause and phrase), we have included one global measure operationalized in terms of length (number of words), one diversity measure (standard deviation of length) and where possible, one specific ratio measure to tap into the various dimensions of syntactic complexity (Norris & Ortega, 2009). Because no tool exists to calculate these measures automatically for French and because manually annotating all the corpus texts was not feasible, we created our own R function to automatically annotate the corpus texts for syntactic units (see Vandeweerd, in press). Table 6 lists the syntactic complexity measures.

Table 6. Measures of Syntactic Complexity

Measure	Definition
Sentence Level	
Mean length of sentence (MLS)	∑ sentence lengths in words / # of sentences
Sentence length diversity (DivS)	Standard deviation of sentence length
T units per sentence (T/S)	# of T-units / # of sentences
T-unit Level	
Mean length of T-unit (MLT)	∑ T-unit lengths in words / # of T-units
T-unit length diversity (DivT)	Standard deviation of t-unit length
Clauses per T-unit (C/T)	# of clauses / # of T-units
Clause Level	
Mean length of clause (MLC)	∑ clause lengths in words / # of clauses
Clause length diversity (DivC)	Standard deviation of clause length
Noun phrases per clause (NP/C)	# of noun phrases / # of clauses
Phrase Level	
Mean length of noun phrase (MLNP)	∑ NP lengths in words / # of noun phrases
Noun phrase length diversity (DivNP)	Standard deviation of NP length

3.2.4. Morphological Complexity

In comparison to English, French has a much more extensive verbal inflection paradigm. As such, learners of French show more continuous development of morphological complexity than learners of English (De Clercq & Housen, 2019). To determine whether this would lead to trade-offs in the development of phraseological complexity, as suggested by Stengers, Boers et al. (2011), we decided to include a measure of morphological complexity in the current study although no such measure was used in Paquot (2018, 2019). Our measure of morphological variation is based on Pallotti's (2015) Morphological Complexity Index (MCI), which measures the diversity of morphological variants of verbs, nouns or adjectives used in a text. We calculated the MCI for verbs only, given the relative richness of the verbal morphological system in written French (see Section 2). The calculation was based on the inflectional information provided by MElt POS tagger (Denis & Sagot, 2012) which tags each verb with with mode, number, person, and tense information. For example, *est* (be.3SG.PRS.IND) is tagged: "m=ind|n=s|p=3|t=pst", corresponding to indicative mode, singular number, third person and present tense. MCI was calculated as the average diversity of these inflectional tags across 100 randomly sampled segments of 10 verbal inflections. The drawback to this method is that it only measures the semantic properties of morphological inflection and does not measure the inflectional affixes directly. For example, both *regarde* (see-3SG.PRS.IND) and *a* (have-3SG.PRS.IND) are treated equally in this analysis despite the fact that they rely on different morphological processes given that the former is a highly regular verb and the latter is an irregular verb. Nonetheless, we feel that this method is valid in that it accounts for the range of morphosemantic processes a learner is able to encode in writing and is not affected by sample-size. This method is also similar to that of Verspoor et al. (2012) as well as Bulté (2013) who counted the number and variety of verbal forms. Bartning and Schlyter's (2004) developmental stages for L2 French are also based on the increasing presence of a variety of morphosyntactic forms. According to Ågren, Granfeldt and Schlyter's (2012: 100), "a learner who uses the present and the perfect tenses to express all events in the past, the present and the future, has a less complex tense/mode/aspect system than a learner who uses the whole range of tense/aspect forms".

3.3. Analysis

Distributions for all variables were first checked for normality by visual inspection of the histograms. This visual inspection revealed that six of the phraseological band-based measures exhibited distributions which were largely skewed towards one highly frequent value (1 or 0). As these were all proportion-based measures, this meant that directly comparing the means for each of these variables was problematic due to ceiling and floor effects. We therefore transformed these variables into binary variables for the bivariate analysis (RQ1).[3] For example, a majority of the texts (63%) in the corpus had a value of 1 for the variable ADVMOD_V_PROP_PMI_NONCOL which represents the proportion of adverbial modifiers in the non-collocation band. This variable was transformed to a binary variable with the following two levels: "ALL" (indicating that all adverbial modifiers in a given text were in the

[3] The untransformed versions of the variables were used in the random forest analysis (RQ2) as random forests are more robust to skewed distributions.

non-collocation band) or "NOT_ALL" (indicating that at least one adverbial modifiers in the text was outside of the non-collocation band).

In order to answer RQ1, we conducted a series of bivariate tests with complexity measures as the dependent variable and proficiency as the independent variable. Chi squared tests were used for all binary phraseological variables except ADVMOD_V_PROP_PMI_HI (the proportion of adverbial modifiers in the high collocation band), which did not meet the assumptions for a chi-squared test (expected values below 5), so a two-sided Fisher's Exact test was used instead. Bonferroni corrections were used to correct for multiple comparisons (0.05/5 =0.01). For the remaining non-binary phraseological variables, the Shapiro-Wilk test of normality and Lavene's test of homogeneity of variance revealed that the assumptions for non-parametric tests could not be met so Kruskall-Wallis tests were used. Pairwise Wilcox Post hoc Tests were used in cases of significant group differences and effect size was calculated according to the formula $r = z/\sqrt{n}$ (Rosenthal, 1994). Bonferroni corrections were again used to correct for multiple comparisons. To compare the three mean-PMI based measures, alpha was set at 0.017 (0.05/3). Likewise, alpha was set at 0.017 for the three RTTR-based measures and 0.008 for the six proportion-based measures (0.05/6). All statistical calculations were conducted in R (R Core Team, 2019).

To answer RQ2, we built a random forest model where the complexity variables and the writing prompts were the predictors and the outcome variable was the global CEFR level of a text. In contrast to Paquot (2018, 2019), regression modelling could not be used because the data did not meet the proportional odds assumption. Untransformed versions of the predictor variables were used in the random forest analysis as random forests are more robust to skewed distributions. The model was built from random samples (n = 26) from each proficiency level because unbalanced classes can skew predictions towards the most frequent class. Descriptive statistics for all complexity measures in the sample are provided in Appendix

I. To ensure that each of the samples were as representative as possible, texts from each topic were sampled according to their relative proportion at each level. For example, the topic of *youth delinquency* made up 40% of the C1 texts, so in the sample of C1 texts (n=26), 40% (=11) of them were from the youth delinquency topic. All variables were then entered into a random forest using the *party* package (Hothorn, Buehlmann, Dudoit, Molinaro, & Van Der Laan, 2006; Strobl, Boulesteix, Kneib, Zeileis, & Achim, 2008; Strobl, Boulesteix, Zeileis, & Hothorn, 2007). To check the accuracy of the model on the non-sampled set, we calculated the accuracy of the model on two alternate samples. This was found to be 0.78 and 0.86 respectively, indicating that the model could also generalize reasonably well over the whole data set. To determine which variables contributed to the classification of the texts, we used the *vip* package (Greenwell, Boehmke, & Gray, 2019) which calculates variable importance by running multiple iterations of the model, each time removing one variable at a time to determine the extent to which removing each variable affects the classification accuracy. Following Levshina (2015), the threshold for variable importance was set as the absolute value of the minimally important variable. In order to determine their effect on the classification, that is to say the assignment of proficiency levels to the texts, partial dependence plots were generated using *pdp* package (Greenwell, 2017) for each of these measures.

4. Results

4.1. Phraseological measures (RQ1)

Adverbial modifier, adjectival modifier and direct object dependencies were extracted from the learner texts and PMI values for each dependency were calculated on the basis of the reference corpus. Dependencies with a high PMI tended to be related to the specific topics elicited by the writing prompts, as in the examples in (1), which relate to linguistic and legal topics. The high

PMI dependencies also tended to be composed of words with very few other collocates (e.g. *atténuant*; ‘attenuating’).

(1)

transcription+phonétique ‘phonetic+transcription’	*locuteur+natif* ‘native+speaker’
circonstance+atténuant ‘attenuating+circumstances’	*purger+peine* ‘serve+sentence’

In addition, some of the high PMI dependencies belonged to specific idioms as in (2):

(2)

bouillir+marmite ‘boil+pot’ to maintain a household	*promettre+monts* ‘promise+mountains’ to make big or unrealistic promises

Furthermore, the high PMI dependencies tended to be composed of relatively low frequency, semantically specific words such as *docilement* (‘meekly’; 0.26/million words). In contrast, the dependencies with lower PMIs (3) tended to belong to more general semantic domains and to be composed of higher frequency words such as *être* (be.INF; 11261/million words). This is particularly evident when looking at the adverbial modifiers. The high PMI dependencies include adverbs of manner that are more semantically specialized (e.g. *soutenir+financièrement;* ‘financially support’) whereas the low PMI dependencies include more general adverbs of degree (*vraiment*; ‘really’), time (*souvent*; ‘often) and negation (*jamais*; ‘never’).

(3)

savoir+plus ‘know more’	*cause+important* ‘important+cause’	*être+sujet* ‘be+subject’

Examples of each of the dependency types extracted from the texts are provided in examples (4) to (6) (PMI in brackets).

(4) **ADVMOD_V dependencies with high PMI:**
punir+sévèrement (10.13), lier+intimement(8.94), coûter+cher(8.86), encadrer+strictement(8), soutenir+financièrement(7.83), frapper+violemment(6.88), étudier+minutieusement(6.73), suivre+docilement(6.38)

ADVMOD_V dependencies with low PMI:
violer+jamais(0.99), être,+indiscutablement(0.99), traîner+souvent(0.98), savoir+plus(0.97), vouloir+vraiment(0.97), disposer+déjà(0.97), manifester+parfois(0.97), dépasser+quelquefois(0.97), tendre+toujours(0.97), opter+souvent(0.97)

(5) **AMOD dependenceis with high PMI :**
démence+sénile(13.76), locuteur+natif(12.24), mineur+délinquant(12.2), délinquance+juvénile(11.62), circonstance+atténuant(11.61), transcription+phonétique(11.21), adulte+consentant(11.15), cercle+vicieux(10.96), alphabet+phonétique(10.24), attentat+terroriste(10.12)
AMOD depenedencies with low PMI :
crime+autre(1), débat+nombreux(0.99), façon+exact(0.99), délit+petit(0.99), cause+important(0.99), matière+complexe(0.99), conduite+ordinaire(0.99), comportement+social(0.99), tranche+différent(0.99), châtiment+léger(0.99)

(6) **DOBJ dependencies with high PMI :**
bouillir+marmite(12.17), légaliser+euthanasie(11.63), dissiper+malentendu(10.61), rectifier+tir(10.06), souffrir+martyr|martyre(9.66), abréger+souffrance(9.41), purger+peine(9.16), graffité+mur(9.16), promettre+mont(9.07), dérailler+train(9.03)
DOBJ dependencies with low PMI :
connaître+valeur(0.98), ressentir+conséquence(0.98), confronter+auteur(0.98), qualifier+thèse(0.97), appliquer+solution(0.97), être+sujet(0.97), adorer+enfant(0.96), faire+apprentissage(0.96), avoir+confiance(0.96), négliger+formation(0.96)

As shown in Table 7 the proportion of texts containing only non-collocating adverbial modifiers decreased from B2 to C2 as did the proportion of texts with at least one low or medium collocating adverbial modifier. The proportion of texts with at least one high collocating adverbial modifier also increased slightly. The proportion of texts with at least one high collocating direct object modifier and adjectival modifier was similar across all three proficiency levels, with a slightly higher proportion of texts with at least one high collocating adjectival modifier in C1 than in B2 or C2. None of these differences were found to be significant.

Table 7. Binary phraseological measures (α = 0.01)

Measure		Levels	B2	C1	C2	Test	χ^2	df	p
ADVMOD_V	NONCOL	ALL:NOT_ALL	13:13	42:64	8:29	CHISQ	5.93	2	0.052
	LOW	ABSENT:PRESENT	15:11	56:50	12:25	CHISQ	5.47	2	0.065
	MED	ABSENT:PRESENT	24:2	82:24	27:10	CHISQ	3.71	2	0.156
	HI	ABSENT:PRESENT	26:0	101:5	33:4	FISHER	NA	2	0.151
DOBJ	HI	ABSENT:PRESENT	17:9	54:52	21:16	CHISQ	1.86	2	0.395
AMOD	HI	ABSENT:PRESENT	9:17	35:71	13:24	CHISQ	0.07	2	0.968

As shown in Table 8, all three mean PMI based sophistication measures showed a linear increase across proficiency levels but this was only significant for direct objects and only between B2-C1 and B2-C2. In the case of the band-based measures, there was a slight increase in the proportion of low and medium collocating adjectival modifiers from B2 to C1 and a decrease in the proportion of non-collocating adjectival modifiers. The proportion of adjectival modifiers remained the same in all bands between C1 and C2. Direct objects showed a different pattern. The proportion in the low and non-collocating bands decreased from B2 to C1 and the proportion in the medium band increased. From C1 to C2, the proportion in the low band increased, the proportion in the medium band decreased and the proportion in the non-collocating band remained constant. The phraseological diversity measures (RTTR) for adjectival and adverbial modifiers showed a U-shaped pattern: decreasing from B2 to C1 but increasing between C1 and C2 (significant only for adjectival modifiers). The diversity of direct-objects showed a significant linear increase across proficiency levels. To summarize, the mean PMI of direct objects and the RTTR of adjectival modifiers increased significantly across proficiency levels. The effect size for each of these are at or below .33 and are considered small (Plonsky & Oswald, 2014).

Table 8. Non-binary phraseological measures (arrows represent the direction of change from one CEFR level to the next)

	Mean (SD)								
Measure	B2		C1		C2	χ^2	α	p	Post hoc
Diversity									
ADVMOD_V_RTTR	4.54(0.84)	↘	4.46(0.87)	↗	4.73(0.67)	3.83	0.017	0.148	
AMOD_RTTR*	4.04(1.1)	↘	3.86(1.06)	↗	4.52(0.97)	9.57	0.017	0.008	C1-C2(r=.21)
DOBJ_RTTR	4.3(0.81)	↗	4.18(0.9)	↗	4.53(0.79)	3.90	0.017	0.142	
Sophistication									
ADVMOD_V_MEAN_PMI	0.29(0.22)	↗	0.41(0.25)	↗	0.44(0.25)	7.03	0.017	0.030	
AMOD_MEAN_PMI	2.3(1.09)	↗	2.67(1.22)	↗	2.78(0.75)	4.95	0.017	0.084	
AMOD_PROP_PMI_NONCOL	0.63(0.16)	↘	0.58(0.18)	→	0.58(0.13)	3.07	0.008	0.215	
AMOD_PROP_PMI_LOW	0.19(0.13)	↗	0.22(0.12)	→	0.22(0.1)	3.07	0.008	0.216	
AMOD_PROP_PMI_MED	0.11(0.09)	↗	0.12(0.12)	→	0.12(0.08)	1.37	0.008	0.504	
DOBJ_MEAN_PMI*	1.51(0.58)	↗	1.9(0.75)	↗	2.01(0.57)	8.99	0.017	0.011	B2-C1(r=.13), B2-C2(r=.33)
DOBJ_PROP_PMI_NONCOL	0.72(0.09)	↘	0.7(0.13)	→	0.7(0.1)	0.84	0.008	0.658	

DOBJ_PROP_PMI_LOW	0.17(0.1)	↘	0.16(0.09)	↗	0.18(0.09)	1.63	0.008	0.444
DOBJ_PROP_PMI_MED	0.08(0.05)	↗	0.11(0.08)	↘	0.1(0.07)	1.65	0.008	0.439

4.2. Random forest model (RQ2)

The random forest model had a classification accuracy of 0.96 which was significantly better (p<.0001) than baseline. As shown by the confusion matrix in Table 9, the model was able to correctly classify most texts. All misclassified texts were at most one level away from correct classification.

Table 9. Confusion matrix for random forest model

		Reference		
		B2	**C1**	**C2**
Predictions	**B2**	26	1	0
	C1	0	24	1
	C2	0	1	25

The lowest variable importance score was -2.96E-3, so only variables with a VI higher than +2.96E-3 were considered important in the model (Levshina, 2015). As shown in Table 10, these include phraseological sophistication measures, lexical diversity and sophistication measures, morphological diversity and syntactic complexity (phrasal) measures. No phraseological diversity measure reached threshold importance. Partial dependence plots for each of these measures are provided in Figure 1. The observed values for each complexity measure are plotted on the x axis. The y axis shows the probability of a text being classified at a given proficiency level (dotted line = B2, dashed line = C1, solid line = C2), as the variable in question increases or decreases and all other variables are held constant. The highest line on the graph therefore represents the proficiency level which is most probable at each value of the complexity measure. These ranges are also provided in Table 10. In general, as these variables increase (or decrease in the case of non-collocating adverbial modifiers), the probability of a text being assigned to a higher level also increases. Three measures show a somewhat curvilinear pattern whereby texts with the lowest values are more likely to be categorized as C1 but texts with mid-range values are more likely to be categorized as B2: VAR_NOUN

(noun variation), VAR_LEX (lexical word variation) and FFD.RTTR.SOPH.NOUN (the proportion of sophisticated nouns). The mid-range values of FFD.RTTR.SOPH.NOUN also have a range where C1 is slightly more likely than B2 to be predicted. Taken together, these results show that compared to B2 texts, C1 texts are more likely to have less noun variation and lexical word variation, but a higher proportion of more sophisticated lexical words (excluding nouns), longer noun phrases and direct object dependencies with a higher mean PMI. In other words, C1 texts tend to be less lexically diverse, but exhibit more lexical and phraseological sophistication and have longer noun phrases. Compared to C1 texts, C2 texts are more likely to have more noun variation and lexical word variation, a higher proportion of sophisticated lexical words (including nouns), fewer non-collocating adverbial modifiers and direct object dependencies with a higher mean PMI. C2 texts are also more likely to have more verb form variation than B2 texts. This means that C2 texts are more likely to be higher in lexical, phraseological and morphological complexity compared to B2 and C1 texts, but not necessarily higher in syntactic complexity at the level of the noun phrase.

Table 10. Prediction ranges according to partial dependency plots

		Prediction Ranges		
Variable	Importance	B2	C1	C2
VAR_NOUN Noun variation (lex. div.)	8.54E-03	0.27-0.28	0.15-0.27	0.29-0.38
MCI.V Morphological Complexity Index (verbs) (morph. div.)	7.50E-03	2.74-5.46	NA	5.58-8.66
FFD.RTTR.SOPH.LEX Proportion of FFD off-list lexical lemmas (lex. soph.)	6.42E-03	0.71-2.49	2.57-3.19	3.27-4.58
DOBJ_MEAN_PMI Mean PMI for DOBJ dependencies (phras. soph.)	6.05E-03	0.42-1.7	1.76-2.08	2.13-3.09
FFD.RTTR.SOPH.NOUN Proportion of FFD off-list noun lemmas (lex. soph.)	4.66E-03	1.94-2.31	0.34-1.87, 2.38-2.53	2.6-3.99
VAR_LEX Lexical word variation (lex. div.)	4.14E-03	0.69-0.72	0.52-0.68	0.72-0.8
ADVMOD_V_PROP_PMI_NONCOL Proportion of non-col ADVMOD_V dependencies (phras. soph.)	3.92E-03	NA	0.97-1	0.87-0.97
MLNP Mean length of noun phrases (synt.)	3.66E-03	3.01-3.37	3.41-5	NA

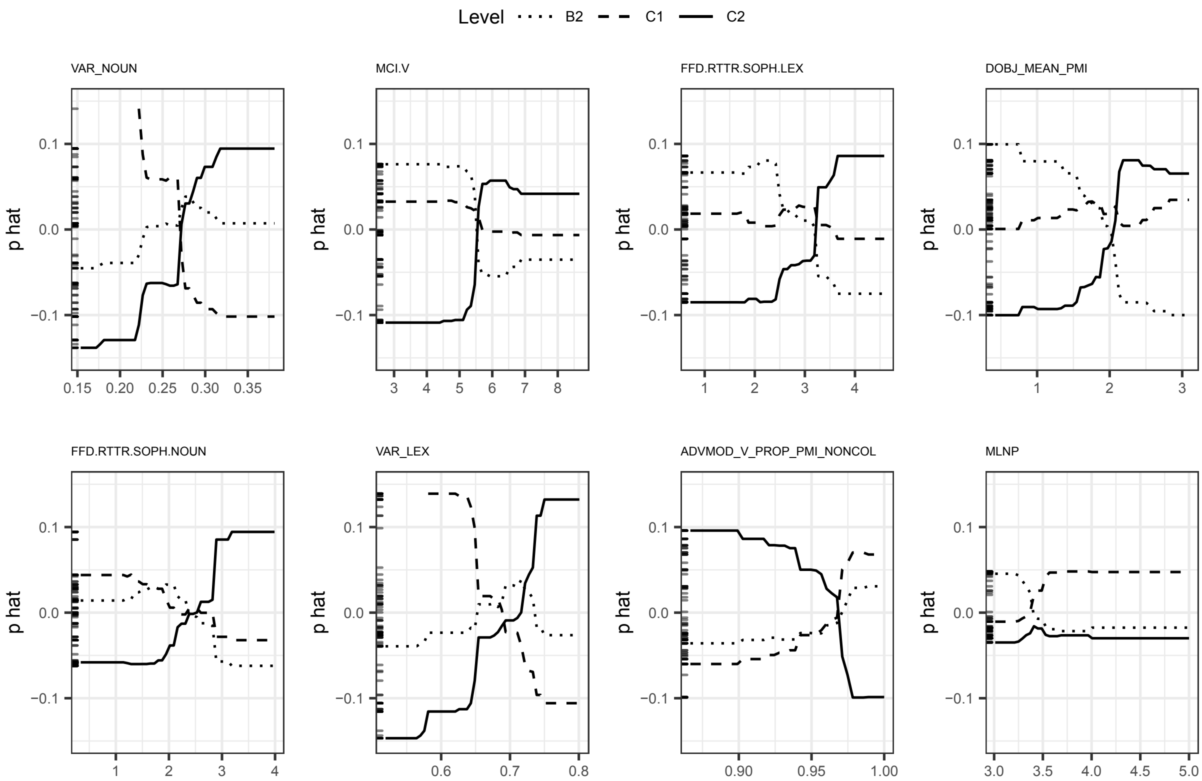

Figure 1. Partial dependency plots for important variables (VI > 2.96E-3)

5. Discussion

5.1. How does phraseological complexity compare in written L2 French at different proficiency levels (RQ1)?

Phraseological diversity was operationalized as the root type-token ratio of adjectival modifier, adverbial modifier and direct object dependencies. Although all three dependency types increased in diversity from B2 to C2, only direct objects showed a linear increase in diversity with proficiency. The diversity of adjectival and adverbial modifiers was lower in C1 texts than B2 texts. Only the increase in diversity for adjectival modifiers between C1 and C2 was found to be significant however. These results are in line with the findings of Erman et al. (2015) who found that unlike L2 English speakers, advanced L2 French speakers did not reach native-like levels of diversity in their use of formulaic expressions. However, they contrast with Paquot's (2019) results for L2 English, which found no systematic increase in phraseological diversity across proficiency levels. But given that the texts in Paquot (2019) were on average 3000 words longer than the texts in the current study and consisted of research papers and not argumentative essays, it is not clear to what extent this difference is due to register differences rather than target language differences. In section 5.2 we discuss this finding further in relation to another domain of complexity, namely morphological complexity.

Phraseological sophistication was operationalized as the mean pointwise mutual information score of dependencies as well as the proportion of dependencies in four collocation bands. Mean PMI was found to increase linearly with proficiency for all three dependencies but the only significant difference was found for direct object dependencies between B2-C1 and B2-C2. Verb-object collocations are known to be difficult for L2 learners in general (see Paquot & Granger, 2012) and these findings are in line with the results of Paquot (2018, 2019) for L2 English and Forsberg Lundell et al. (2018) for L2 French. It is worth noting that where Paquot (2018, 2019) observed a similar range for mean PMI across all three dependency types, in the L2 French data, the mean PMI for adverbial modifiers is much lower than that of adjectival modifiers and direct objects. This suggests that even at very advanced levels of L2 French, the association strength of adverbial modifier dependency relations remains relatively low compared to adjectival modifiers and direct objects. This may be due to the stylistic preferences of French, which unlike English, tends to use prepositional phrases to modify verbs rather than adverbs (Vinay & Darbelnet, 1995). For example, the idiomatic French equivalent of “answered angrily” is *répondu avec colère* (‘answered with anger’). This type of verbal modification is not captured by the dependency-based measures used in the current study. With respect to the proportion-based measures, although no significant differences were found for these measures across proficiency levels, there was a general trend whereby the proportion of non-collocating dependencies decreased with proficiency and the proportion of high-collocating dependencies remained constant. The proportion of low and medium collocating dependencies tended to increase with proficiency, consistent with the results of Paquot (2018) for L2 English.

In general, from B2 to C1, there was an increase in phraseological sophistication and a decrease in phraseological diversity. From C1 to C2, there was an increase in both phraseological diversity and sophistication. However, significant differences between proficiency levels were only found for the diversity of adjectival modifiers between C1 and C2 and the sophistication of direct objects between the B and C levels.

5.2. To what extent does phraseological complexity relate to lexical, syntactic and morphological complexity in L2 written French (RQ2)?

The results of the random forest analysis showed that all four domains of complexity (lexical, syntactic, morphological and phraseological) contributed to predictions of the proficiency levels of the texts in the corpus. The most important predictors of proficiency level included measures of both lexical diversity (noun variation and lexical word variation) and lexical sophistication (proportion of sophisticated nouns and sophisticated lexical words), in line with previous research linking L2 French proficiency with lexical diversity (De Clercq, 2015) and sophistication (Lindqvist, Gudmundson, & Bardel, 2013; Ovtcharov et al., 2006). Though the predictors for the model included lexical sophistication measures composed of both absent words from the French frequency dictionary as well as on-list words from the Lexique Transdisciplinaire list, only the former were found to be significant predictors in the model. This might provide some evidence for Cobb and Horst's (2004) claim that an academic word list may not be necessary to capture lexical sophistication in formal written L2 French, at least not for the argumentative essay register used in the current study. Morphological diversity (MCI for verbs) was also an important predictor in the model, consistent with previous research showing significant proficiency differences for this measure as well as higher levels of morphological diversity in learners of French compared to learners of English (De Clercq, 2016; De Clercq & Housen, 2019). The fact that morphological diversity continued to play a role in the classification of advanced texts suggests that the impact of phraseological complexity on proficiency may indeed be influenced by morphology, as suggested by Stengers et al. (2011). Only one measure of syntactic complexity was an important predictor of proficiency (mean length of noun phrases). The only L2 French study to measure syntactic complexity at the phrasal level, De Clercq and Housen (2017), did not find a significant difference between beginner-intermediate proficiency levels for this measure, but did find a significant difference between the learners and a native-speaker control, which provides evidence that this measure may be more discriminatory at the advanced level. That complexity measures in all four domains were important predictors in the model contrasts with Paquot's (2018, 2019) results for L2 English, whose final model contained only phraseological measures. Again, the differences between this study and Paquot (2018, 2019) regarding text length and register means that the current results should be interpreted with caution. That being said, these results are consistent with those of De Clercq (2016) to the extent that lexical diversity was also an important predictor for all three proficiency levels. Furthermore, these results show that syntactic elaboration at the level of the noun phrase is the most important syntactic complexity measure for distinguishing between intermediate and advanced learners, again consistent with De Clercq's (2016) finding that this measure distinguished between high school learners and native speakers. Unlike De Clercq (2016) however, lexical sophistication and morphological diversity also seem to play a role in distinguishing the most advanced proficiency levels, at least in the register of argumentative writing used in the current study. As in Paquot (2019), measures of phraseological diversity (root type-token ratio of dependencies) were not important predictors of proficiency.[4] At first glance, it may be surprising that the diversity of

[4] A recent study of L2 Dutch found that the diversity of adverbial modifiers *was* a significant predictor of passing proficiency exams (Rubin, Housen, & Paquot, in press). However, the effect of this predictor was strongest in the B1 exams and much weaker in the B2 exams, which may explain why this measure was not found to be

adjectival modifiers was not an important predictor in the random forest model, given that the diversity of adjectival modifiers was found to be significantly higher in C2 texts compared to C1 texts. What this seems to indicate is that when all other variables are controlled for (including the diversity of nouns and adjectives), the diversity of adjectival modifiers is no longer an important predictor in the model. In other words, the variation of lexical words (nouns, verbs, adjectives and adverbs together) and nouns (separately) accounts for the diversity of adjective + noun combinations. This indeed seems to be the case when looking at the correlations between these two measures. The diversity of adjectival modifiers is highly and significantly correlated with the diversity of adjectives ($r = 0.89$, $p < 0.001$) as well as the diversity of nouns ($r = 0.75$, $p < 0.001$). Phraseological sophistication on the other hand, cannot be accounted for by lexical sophistication measures alone because even when lexical sophistication is controlled for, two phraseological sophistication measures are still important to the model (the mean PMI of direct objects and the proportion of non-collocating adverbial modifiers). In contrast to the strong correlation between phraseological and lexical diversity measures, the mean PMI of direct objects is not significantly correlated with sophistication of nouns ($r = 0.01$, $p = 0.91$) nor the sophistication of verbs ($r = -0.20$, $p = 0.08$). Likewise, the proportion of non-collocating adverbial modifiers is not significantly correlated to the sophistication of adverbs ($r = -0.05$, $p = 0.67$) nor the sophistication of verbs ($r = -0.22$, $p = 0.05$). In other words, whereas phraseological diversity patterns closely with measures of lexical diversity, the same does not seem to be true for phraseological sophistication and lexical sophistication.

That the mean PMI of direct objects and the proportion of non-collocating adverbial modifiers are important predictors of proficiency level, even when controlling for the diversity

significant in Paquot's (2019) study or in this study of L2 French, which both focused on learners beyond the B1 level.

of verbal inflections, indicates that these two measures are useful indices of proficiency in a synthetic language such as French. However, the possibility that morphological processes somewhat dampen the contribution of phraseological complexity cannot be ruled out either, given that morphological diversity was found to be an important predictor of proficiency, even at the advanced level.

6. Conclusion

The main aim of this paper was to determine how measures of phraseological complexity compare across proficiency levels in L2 French and to determine how these measures compare to other measures of lexical, syntactic and morphological complexity. As a partial replication study, the goal was also to see whether the phraseological complexity measures which were originally developed by Paquot (2018, 2019) for L2 English, would also be predictive of proficiency in L2 French.

As in Paquot (2018, 2019), phraseological complexity was operationalized as the diversity and sophistication of adjectival modifier, adverbial modifier and direct object dependencies, which were automatically extracted from a corpus of L2 French argumentative texts using a dependency parser. In addition to the phraseological measures, measures of lexical, syntactic and morphological complexity were also calculated. Though all phraseological complexity measures increased with proficiency, this was only significant for the diversity of adjectival modifiers between C1-C2 and the sophistication of direct objects between B2-C1 and B2-C2. A random forest model based on the complexity measures was found to have a high level of accuracy in classifying the texts according to the holistic proficiency levels they were assigned by trained CEFR raters. As in Paquot (2019), phraseological sophistication measures were shown to be important predictors in the model, which seems to indicate the usefulness of phraseological sophistication as an index of advanced L2 French proficiency.

In contrast to Paquot (2018, 2019), who found that phraseological measures were better predictors of proficiency than traditional complexity measures, the current study found that the most important predictors in the model also included measures of complexity in other linguistic domains: namely lexical diversity, lexical sophistication, morphological diversity and syntactic elaboration (phrasal). However, because this is a replication rather than a direct cross-linguistic comparison, one must be cautious when interpreting any differences between the current study and Paquot (2018, 2019). Although every attempt was made to replicate the methods and the dataset, the limited availability of data and tools for L2 French did not always make this possible. The learner corpus of the current study was shorter in mean text length and was composed of argumentative essays instead of research papers. Given the small-scale of the learner corpus used, it was also not possible to split the data into a test set and an evaluation set for model building and as such, the generalizability of these results to other data sets will need to be evaluated in future studies. It may also be fruitful in future research to focus explicitly on cross-linguistic comparisons of phraseological complexity but as the findings of the current study suggest, it will be important to carefully control for the effect of morphology on phraseological complexity as well as the stylistic preferences, for example prepositional modifiers (e.g. répondre avec colère; 'respond with anger') instead of adverbial modifiers.

The current study is part of a larger research project and so efforts to replicate the results with other L2 French corpora are currently underway. That being said, the results of the current study are consistent with the (limited) existing literature on L2 French complexity and, along with Paquot (2018, 2019), speak to the importance of including phraseology in the current repertoire of L2 complexity measures for the purpose of assessing L2 proficiency.

Supplementary Materials

The R scripts that were used to calculate phraseological, lexical and morphological complexity measures, as well as the scripts and data for the statistical analysis are available here: https://osf.io/mwyg6. Also see Vandeweerd (in press) for the function that was used to extract syntactic units.

Acknowledgement

We would like to thank Hubert Naets for his invaluable help extracting the grammatical dependencies and the calculation of PMI. We are also very grateful to Anthonya Delfosse and Héloïse Copin for their help with the annotations. The lexical complexity scripts were adapted from scripts written by Vanhove (2018). This work was supported by the Fonds de la Recherche Scientifique – FNRS under Grant n° T.0086.18

References

Ågren, M., Granfeldt, J., & Schlyter, S. (2012). The growth of complexity and accuracy in L2 French: Past observations and recent applications of developmental stages. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 95–120). Amsterdam: John Benjamins.

Bartning, I., & Schlyter, S. (2004). Itinéraires acquisitionnels et stades de développement en français L2 [Acquisitional routes and stages of development in L2 French]. *Journal of French Language Studies*, *14*, 281–299. https://doi.org/10.1017/S0959269504001802

Batista, R., & Horst, M. (2016). A new receptive vocabulary size test for French. *Canadian Modern Language Review*, *72*(2), 211–233. https://doi.org/10.3138/cmlr.2820

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*. Leiden: Brill Rodopi.

Blanche-Benveniste, C., & Adam, J.-P. (1999). La conjugaison des verbes: Virtuelle, attestée, defective [The conjugation of verbs: Virtual, attested, defective]. *Recherches Sur Le Français Parlé*, *15*, 87–112.

Bulté, B. (2013). *The development of complexity in second language acquisition: A dynamic systems approach*. (Unpublished doctoral dissertation). Vrije Universiteit Brussel, Brussels, Belgium.

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). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.32.02bul

Candito, M., Nivre, J., Denis, P., & Anguiano, E. H. (2010). Benchmarking of statistical dependency parsers for French. In *COLING 2010: Poster Volume* (pp. 108–116). Beijing.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography, *16*, 22–29. https://doi.org/10.3115/981623.981633

Cobb, T., & Horst, M. (2004). Is there room for an academic word list in French? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language : Selection, acquisition, and testing* (pp. 15–38). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.10.04cob

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

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238.

https://doi.org/10.2307/3587951
De Clercq, B. (2015). The development of lexical complexity in second language acquisition: A cross-linguistic study of L2 French and English. *EUROSLA Yearbook*, *15*, 69–94. https://doi.org/10.1075/eurosla.15.03dec
De Clercq, B. (2016). *The development of linguistic complexity*. (Unpublished doctoral dissertation). Vrije Universiteit Brussel, Brussels, Belgium.
De Clercq, B., & Housen, A. (2017). A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. *Modern Language Journal*, *101*(2), 315–334. https://doi.org/10.1111/modl.12396
De Clercq, B., & Housen, A. (2019). The development of morphological complexity: A cross-linguistic study of L2 French and English. *Second Language Research*, *35*(1), 71–97.
Demol, A., & Hadermann, P. (2008). An exploratory study of discourse organisation in French L1, Dutch L1, French L2 and Dutch L2 written narratives. In *Linking up contrastive and learner corpus research* (pp. 255–282). Amsterdam: Rodopi. https://doi.org/10.1163/9789401206204_011
Denis, P., & Sagot, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, *46*(4), 721–736. https://doi.org/10.1007/s10579-012-9193-0
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. https://doi.org/10.1515/iral.2009.007
Erman, B., Denke, A., Fant, L., & Forsberg Lundell, F. (2015). Nativelike expression in the speech of long-residency L2 users: A study of multiword structures in L2 English, French and Spanish. *International Journal of Applied Linguistics*, *25*(2), 160–182. https://doi.org/10.1111/ijal.12061
Forsberg, F., & Bartning, I. (2010). Can linguistic features discriminate between the communicative CEFR-levels? : A pilot study of written L2 French. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development : Intersections between SLA and language testing* (pp. 133–157). European Second Language Association.
Forsberg Lundell, F., Lindqvist, C., & Edmonds, A. (2018). Productive collocation knowledge at advanced CEFR levels: Evidence from the development of a test for advanced L2 French. *Canadian Modern Language Review*, *74*(4), 627–649. https://doi.org/10.3138/cmlr.2017-0093
Garner, J., Crossley, S. A., & Kyle, K. (2018). N-gram measures and L2 writing proficiency. *System*, *80*, 176–187. https://doi.org/10.1016/j.system.2018.12.001
Goldberg, A. E. (2006). *Constructions at work*. Oxford: Oxford University Press.
Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, *52*(3), 229–252. https://doi.org/10.1515/iral-2014-0011
Greenwell, B. (2017). pdp: An R package for constructing partial dependence plots. *The R Journal*, *9*(1), 421–436.
Greenwell, B., Boehmke, B., & Gray, B. (2019). vip: Variable importance plots. Retrieved from https://cran.r-project.org/package=vip
Guiraud, P. (1954). *Les charactères statistiques du vocabulaire : Essai de méthodologie [The statistical characteristics of vocabulary: Methodological test]*. Paris: Presses Universitaires de France.
Gyllstad, H., Granfeldt, J., Bernardini, P., & Källkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in

written L2 English, L3 French and L4 Italian. *EUROSLA Yearbook*, *14*, 1–30. https://doi.org/10.1075/eurosla.14.01gyl
Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. (2006). Survival Ensembles. *Biostatistics*, *7*(3), 355–373.
Hunston, S., & Francis, G. (2000). *Pattern Grammar*. Amsterdam: John Benjamins.
Levshina, N. (2015). Conditional inference trees and random forests. In *How to do linguistics with R: Data exploration and statistical analysis* (pp. 291–300). Amsterdam: John Benjamins. https://doi.org/10.1075/z.195
Lindqvist, C., Gudmundson, A., & Bardel, C. (2013). A new approach to measuring lexical sophistication in L2 oral production. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis (Eurosla Monograph Series)* (Vol. 2, pp. 109–126). EUROSLA.
Lonsdale, D., & Le Bras, Y. (2009). *A Frequency Dictionary of French*. New York: Routledge.
Michalke, M. (2019). koRpus: An R Package for Text Analysis (Version 0.12-1).
Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555–578. https://doi.org/10.1093/applin/amp044
Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Szmrecsanyi & B. Kortmann (Eds.), *Linguistic complexity: Second language acquisition, indigenization, contact* (pp. 127–155). Berlin: de Gruyter.
Ovtcharov, V., Cobb, T., & Halter, R. (2006). La richesse lexicale des productions orales: Mesure fiable du niveau de compétence langagière. *Canadian Modern Language Review*, *63*(1), 107–125. https://doi.org/10.3138/cmlr.63.1.107
Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*(11), 117–134. https://doi.org/10.1177/0267658314536435
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*(1), 29–43. https://doi.org/10.1080/15434303.2017.1405421
Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145. https://doi.org/10.1177/0267658317694221
Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*(2012), 130–149. https://doi.org/10.1017/S0267190512000098
Peters, E., Velghe, T., & Van Rompaey, T. (2019). The VocabLab tests: The development of an English and French vocabulary test. *International Journal of Applied Linguistics*, *170*(1), 53–78. https://doi.org/10.1075/itl.17029.pet
Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079
Porte, G. (2012). *Replication research in applied linguistics*. Cambridge: Cambridge University Press.
R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from https://www.r-project.org/
Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.
Rubin, R., Housen, A., & Paquot, M. (in press). Phraseological complexity as an index of L2 Dutch writing proficiency: A partial replication study. In *Perspectives on the Second Language Phrasicon: The View from Learner Corpora*.

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)*. Lancaster: IDS.

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, … S. Piperidis (Eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 486–493). Istanbul, Turkey: European Language Resources Association (ELRA).

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532. https://doi.org/10.1093/applin/amp047

Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, *33*(2), 149–183. https://doi.org/10.1177/0741088316631527

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 in Language Teaching*, *49*, 321–343. https://doi.org/10.1515/iral.2011.017

Strobl, C., Boulesteix, A.-L., Kneib, T., Zeileis, T., & Achim, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(307).

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(25).

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79–104). Amsterdam/Philidelphia: John Benjamins.

Tutin, A., & Grossman, F. (2014). *L'écrit scientifique : du lexique au discours. Autour de Scientext [Scientific writing: from lexis to discourse. Overview of Scientext]*. Rennes: Presses Universitaires de Rennes.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Vanderbauwhede, G. (2012). *Le déterminant démonstratif en français et en néerlandais à travers les corpus : Théorie, description, acquisition [The demonstrative determiner in French and Dutch corpora: Theory, description and acquisition]*. (Unpublished Doctoral Dissertation). Katholieke Universiteit Leuven and Université Paris Ouest Nanterre La Défense.

Vandeweerd, N. (in press). synt.units.fcn(): An R function to extract syntactic units from L2 French texts. *International Journal of Learner Corpus Research*.

Vanhove, J. (2018). Computer code for cleaning, tagging, and analysing the texts. Retrieved from https://osf.io/479um/

Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, *21*(3), 239–263. https://doi.org/10.1016/j.jslw.2012.03.007

Vinay, J.-P., & Darbelnet, J. (1995). *Comparative stylistics of French and English*. (J. Sager & M.-J. Hamel, Trans.). Amsterdam: John Benjamins.

Welcomme, A. (2013). Jonction interpropositionnelle et complexité syntaxique dans les récits d'apprenants néerlandophones et locuteurs natifs du français [Clausal linking and syntactic complexity in the writing of Dutch learners and French native speakers]. In U. Paprocka-Piotrowska, C. Martinot, & S. Gerolimich (Eds.), *La complexité en langue et son acquisition* (pp. 261–284). Lublin: Towarzystwo Naukowe Kul Katolicki Uniwersytet Jana Pawla II.

Appendix I: Descriptive statistics for sampled texts (n=78)

		Mean (SD)		
	Variable	B2 (n=26)	C1 (n=26)	C2 (n=26)
Lexical Diversity	RTTR	10.2(1.24)	10.04(1.39)	10.84(1.06)
	RTTR_ADJ*	4.99(1.06)	4.91(1.17)	5.3(0.81)
	RTTR_ADV*	3.68(0.46)	3.57(0.56)	3.85(0.53)
	RTTR_NOUN*	6.66(1.29)	6.28(1.6)	7.33(1.3)
	RTTR_VERB*	6.27(0.97)	5.95(1.2)	6.55(1)
	VAR_LEX	0.68(0.05)	0.65(0.06)	0.7(0.05)
	VAR_ADJ	0.11(0.02)	0.11(0.03)	0.12(0.02)
	VAR_ADV	0.09(0.02)	0.09(0.02)	0.09(0.02)
	VAR_MOD	0.2(0.03)	0.2(0.03)	0.21(0.03)
	VAR_NOUN	0.26(0.04)	0.24(0.05)	0.28(0.04)
	VAR_VERB	0.21(0.04)	0.2(0.03)	0.22(0.02)
Lexical Sophistication	FFD.RTTR.SOPH.ADJ	1.47(0.77)	1.66(0.61)	1.87(0.52)
	FFD.RTTR.SOPH.ADV	0.54(0.28)	0.54(0.37)	0.58(0.45)
	FFD.RTTR.SOPH.LEX	2.48(0.71)	2.49(0.91)	3.03(0.92)
	FFD.RTTR.SOPH.NOUN	1.83(0.5)	1.7(0.84)	2.22(0.9)
	FFD.RTTR.SOPH.VERB	0.98(0.4)	0.99(0.6)	1.2(0.51)
	ADJ.PROP.ONLIST.LEXTRANS*	0.4(0.1)	0.41(0.1)	0.37(0.09)
	N.PROP.ONLIST.LEXTRANS*	0.34(0.06)	0.37(0.07)	0.36(0.07)
	V.PROP.ONLIST.LEXTRANS*	0.81(0.06)	0.81(0.08)	0.79(0.07)
Morphological Diversity	MCI.V*	5.64(1.41)	5.61(1.24)	6.3(0.92)
Syntactic Complexity	MLS*	18.04(3.79)	19.06(3.91)	19.01(2.39)
	DIVS*	8.25(2.23)	8.27(2.23)	8.33(2.12)
	T_S*	1.12(0.12)	1.11(0.09)	1.11(0.09)
	MLT*	15.77(2.67)	17.16(3.44)	16.98(2.18)
	DIVT*	7.53(2.06)	7.75(2.26)	7.91(2.01)
	C_T	1.45(0.21)	1.64(0.42)	1.53(0.3)
	MLC	14.07(1.78)	14.64(1.89)	14.78(1.81)
	DIVC*	8.11(2.15)	8.48(1.83)	8.3(1.75)
	MLNP*	3.44(0.29)	3.63(0.39)	3.54(0.4)
	DIVNP*	2.8(0.9)	3.08(1)	2.86(1.13)
	NP_C	2.7(0.7)	2.62(0.62)	2.93(1.05)
Phraseological Diversity	ADVMOD_V_RTTR	4.54(0.84)	4.51(0.78)	4.86(0.72)
	AMOD_RTTR	4.04(1.1)	3.98(1.07)	4.46(0.91)
	DOBJ_RTTR	4.3(0.81)	4.28(0.89)	4.5(0.86)
Phraseological Sophistication	ADVMOD_V_MEAN_PMI	0.29(0.22)	0.4(0.28)	0.44(0.25)
	ADVMOD_V_PROP_PMI_HI	0(0)	0(0.01)	0(0.01)
	ADVMOD_V_PROP_PMI_LOW	0.02(0.03)	0.01(0.02)	0.03(0.03)
	ADVMOD_V_PROP_PMI_MED	0(0.01)	0.01(0.02)	0.01(0.02)
	ADVMOD_V_PROP_PMI_NONCOL	0.98(0.03)	0.98(0.03)	0.95(0.04)
	AMOD_MEAN_PMI	2.3(1.09)	2.65(1.01)	2.7(0.69)
	AMOD_PROP_PMI_HI	0.08(0.08)	0.07(0.09)	0.08(0.07)
	AMOD_PROP_PMI_LOW	0.19(0.13)	0.21(0.11)	0.21(0.09)
	AMOD_PROP_PMI_MED	0.11(0.09)	0.13(0.12)	0.12(0.07)
	AMOD_PROP_PMI_NONCOL	0.63(0.16)	0.59(0.15)	0.59(0.12)
	DOBJ_MEAN_PMI	1.51(0.58)	1.91(0.62)	1.97(0.6)
	DOBJ_PROP_PMI_HI	0.03(0.04)	0.03(0.03)	0.03(0.04)
	DOBJ_PROP_PMI_LOW	0.17(0.1)	0.18(0.09)	0.17(0.1)
	DOBJ_PROP_PMI_MED	0.08(0.05)	0.11(0.08)	0.1(0.08)
	DOBJ_PROP_PMI_NONCOL	0.72(0.09)	0.68(0.1)	0.69(0.1)

*measures which were not in Paquot (2018, 2019)