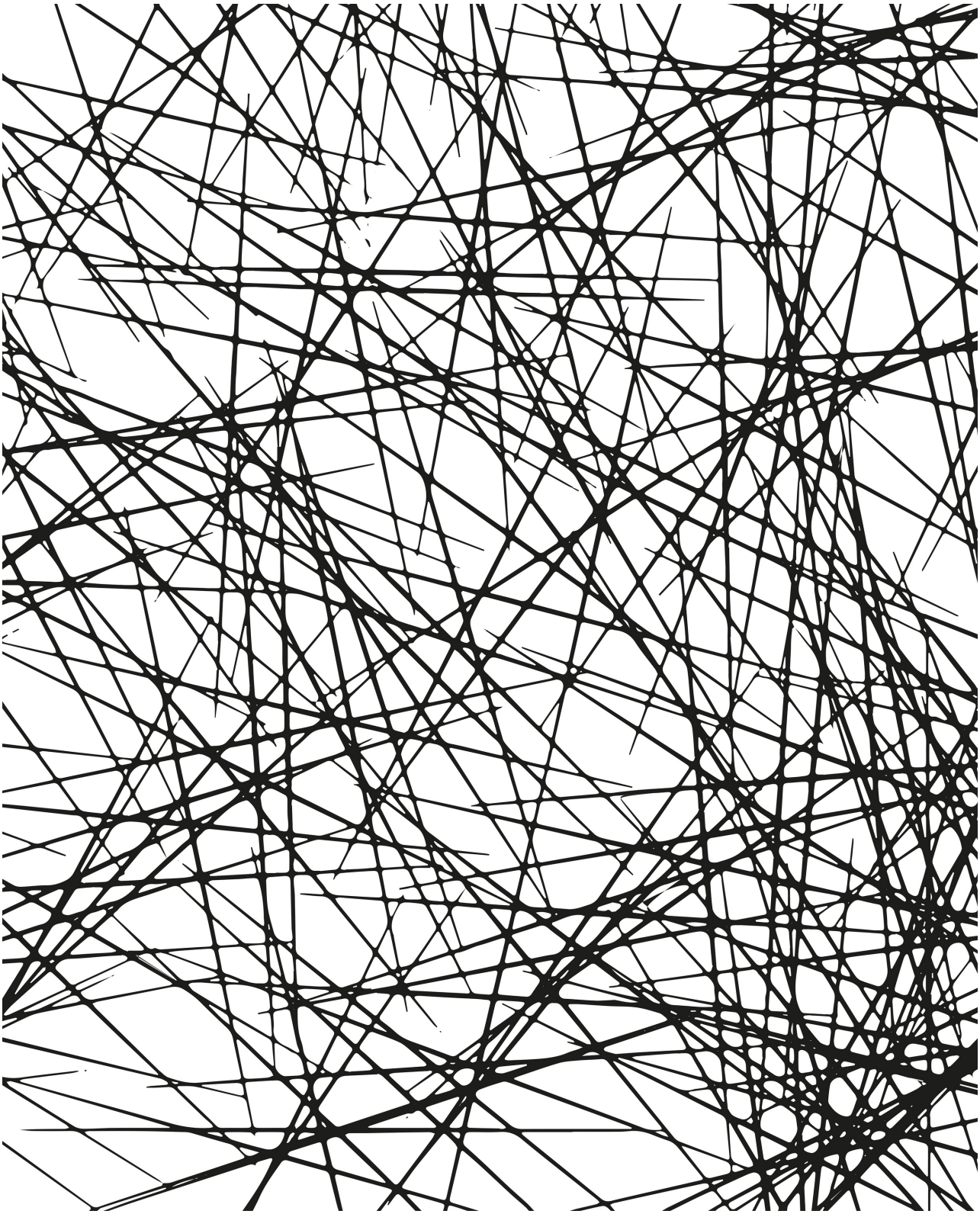
Phraseological complexity in L2 French:

Investigating variation across modes

Nathan Vandeweerd



UCLouvain

Faculté de philosophie, arts et lettres
Commission doctorale du domaine « Langues, lettres et traductologie »
IL&C – Centre for English Corpus Linguistics

VUB VRIJE UNIVERSITEIT BRUSSEL

Faculteit Letteren en Wijsbegeerte
Doctoral School of Human Sciences
Center for Linguistics

Phraseological complexity in L2 French:

Investigating variation across modes

Nathan Vandeweerd

Dissertation submitted for the joint degree of
Docteur en Langues, lettres et traductologie, UCLouvain
Doctor in de Taalkunde, Vrije Universiteit Brussel

Examination Committee :

Chair
Dr. Thomas François — UCLouvain

Supervisors
Dr. Magali Paquot — UCLouvain
Prof. Dr. Alex Housen — Vrije Universiteit Brussel

Jury Members
Prof. Douglas Biber — Northern Arizona University
Dr. Bastien De Clercq — Vrije Universiteit Brussel
Prof. Fanny Forsberg Lundell — Stockholm University
Prof. Dr. Marije Michel — University of Groningen

Louvain-la-Neuve
Academic year 2021-2022

This work is part a larger project entitled *Lexicogrammatical complexity: a promising construct for second language research*, which aims to define and circumscribe the linguistic construct of lexicogrammatical complexity and to theoretically and empirically demonstrate its relevance for L2 complexity research and theories of L2 use and development. 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.

Contents

List of Tables

List of Figures

Summaries

English abstract

Along with accuracy and fluency, complexity is considered a major component of L2 performance and proficiency (Housen et al., 2012; Housen & Kuiken, 2009; Skehan, 1996, 2009b). Until recently, however, most measures of complexity have focused on solely lexical or syntactic aspects of L2 production, disregarding the important role that word combinations play in the development of linguistic competence. The construct of *phraseological complexity* (Paquot, 2019) addresses this by measuring complexity at the interface of lexis and grammar. In various studies, measures of phraseological diversity (e.g., type-token ratios of phraseological units) and sophistication (e.g., average PMI of phraseological units) have been found to be useful indices of L2 proficiency (e.g., Jiang et al., 2021; Paquot, 2018, 2019; Rubin et al., 2021) and development (e.g., Bestgen & Granger, 2018; Siyanova-Chanturia, 2015). With a few exceptions however, research in this domain has focused predominantly on L2 English, and so relatively little is known about the applicability of phraseological complexity measures to more synthetic languages such as French. In addition, while a handful of studies have investigated phraseological complexity measures in the oral mode (e.g., Paquot et al., in press), there has not yet been an attempt to compare phraseological complexity across modes. This is important because modality may affect both the type of phraseological units used (Biber et al., 2004) as well as the ability to pay attention to linguistic output, which is hypothesized to lead to generally higher levels of complexity in writing as compared to speaking (Kormos, 2014;

Manchón, 2014; Skehan, 2014).

While research in L2 French has shown that learners use a higher quantity of phraseological units as they increase in proficiency (e.g., Forsberg, 2010), no study thus far has compared phraseological complexity (in terms of diversity and sophistication) across proficiency levels in L2 French or examined the effect of mode (speech vs. writing) on the manifestation of phraseological complexity. This project fills the gap by investigating the extent to which phraseological complexity is an indicator of proficiency and development in L2 French (RQ1), the extent to which phraseological complexity measures relate to other aspects of linguistic complexity (i.e., syntactic, lexical and morphological) in L2 French (RQ2) and the extent to which phraseological complexity differs in oral versus written L2 production (RQ3).

These questions are addressed in a series of empirical studies using both cross-sectional and longitudinal corpora of matched oral and written tasks by the same learners. Overall, the results show that in L2 French, both phraseological and non-phraseological complexity measures are significant predictors of proficiency level (as established independently on the basis of expert raters' holistic proficiency assessments) and that phraseological complexity develops gradually over time. However, in both cases, these effects are small and appear to be moderated by both communicative function and the constraints of online production in speech. These findings speak to the importance of a more 'organic' approach to complexity (Norris & Ortega, 2009) that takes into account the effect of aspects such as target language or modality when interpreting the meaning of complexity measures.

Résumé en français

Tout comme la précision et la fluidité, la complexité est considérée comme une composante majeure de la performance et de la compétence des apprenants L2 (Housen et al., 2012; Housen & Kuiken, 2009; Skehan, 1996, 2009b). Jusqu'à récemment, cependant, la plupart des mesures de complexité ont privilégié les aspects lexicaux ou syntaxiques d'un texte, sans tenir compte du rôle important que jouent les combinaisons de mots dans le développement de la compétence langagière. La *complexité phraséologique* (Paquot, 2019) répond à ce besoin en mettant en lumière la complexité à l'interface du lexique et de la grammaire. Plusieurs études ont montré que les mesures de diversité (p. ex., le rapport entre types et tokens des unités phraséologiques) et de sophistication (p. ex., le PMI moyen des unités phraséologiques) constituent de bons indicateurs de la compétence (Jiang et al., 2021; Paquot, 2018, 2019; Rubin et al., 2021) et du développement (Bestgen & Granger, 2018; Siyanova-Chanturia, 2015) en L2. Or, à quelques exceptions près, ces études ont principalement porté sur l'anglais L2, et on sait très peu de choses sur les langues plus synthétiques comme le français. En outre, très peu d'études ont utilisé des données orales (cf. Paquot, sous presse), et aucune étude n'a pas comparé la complexité phraséologique à travers des deux modalités. Ceci est très important dans la mesure où les unités phraséologiques utilisées ne sont pas les mêmes à l'oral et à l'écrit (Biber et al., 2004). On suppose également que l'apprenant puisse prêter plus d'attention à la production écrite, de sorte que l'on observe, en générale, un niveau de complexité plus élevé dans celle-ci par rapport à la production orale (Kormos, 2014; Manchón, 2014; Skehan, 2014).

Les études sur le français L2 ont montré une plus grande quantité d'unités phraséologiques chez les apprenants plus avancés (p. ex., Forsberg, 2010), mais jusqu'à présent personne n'a comparé la complexité phraséologique (en termes de diversité et sophistication) à travers de différents niveaux de compétence en français L2. L'effet de modalité (l'oral ou l'écrit) sur la complexité phraséologique reste aussi une lacune à combler. À cet égard, ce projet vise à déterminer dans quelle mesure la complexité phraséologique est un indicateur

de compétence et développement en français L2 (QdR1), dans quelle mesure la complexité phraséologique est liée à d'autres aspects de la complexité linguistique (au niveau syntaxique, lexical et morphologique) en français L2 (QdR2), et finalement, dans quelle mesure la complexité phraséologique diffère entre la production orale et écrite des apprenants L2 (QdR3).

Cette thèse comprend trois études empiriques basées sur des corpus d'apprenants, et faisant recours à des données transversales et longitudinales ainsi qu'à des données orales et écrites produites par les mêmes apprenants. Dans l'ensemble, les résultats montrent qu'en français L2, les mesures de complexité phraséologique et non-phraséologique peuvent prédire de façon significative le niveau de compétence d'un apprenant (déterminé sur base d'évaluations holistiques attribuées aux productions par des experts) et que la complexité phraséologique s'améliore progressivement au fil du temps. Cela dit, dans les deux cas, les effets sont faibles et semblent d'être modérés à la fois par la fonction communicative et par les contraintes de production en temps réelle pour la modalité orale. Ces résultats soulignent l'importance d'une approche plus 'organique' à la complexité (Norris & Ortega, 2009), qui prend en compte des aspects tels que la langue cible et la modalité lors de l'interprétation des mesures de complexité.

Nederlandse samenvatting

Complexiteit wordt, tezamen met accuraatheid en vlotheid, gezien als een belangrijke component van T2 prestatie en vaardigheid (Housen et al., 2012; Housen & Kuiken, 2009; Skehan, 1996, 2009b). Echter werden tot voor kort alleen lexicale en syntactische aspecten als maatstaven gebruikt, waardoor de belangrijke rol van woordcombinaties werd genegeerd. Het concept *fraseologische complexiteit* (Paquot, 2019) houdt juist wel rekening met woordcombinaties en meet complexiteit op het grensvlak met lexis en grammatica. Verscheidene studies hebben dan ook aangetoond dat fraseologische diversiteit (bijv. type/token ratio (TTR) van fraseologische eenheden) en verfijndheid (sophistication; bijv. gemiddelde puntsgewijze onderlinge informatie (PMI) van fraseologische eenheden) belangrijke graatmeters zijn voor T2 vaardigheid (bijv.: Jiang et al., 2021; Paquot, 2018, 2019; Rubin et al., 2021) en ontwikkeling (bijv.: Bestgen & Granger, 2018; Siyanova-Chanturia, 2015). Op een aantal uitzonderingen na, is het merendeel van de studies in dit werkveld verricht op T2 Engels, waardoor er maar weinig bekend is over de toepasbaarheid van fraseologische complexiteitsmaten op een meer synthetische taal, zoals Frans. Alhoewel een aantal studies zich focusten op de mondelinge modus (e.g. Paquot et al., in press), is er nog geen studie geweest die fraseologische complexitiet heeft gemeten in meerdere modi. Dit is belangrijk omdat modaliteit zowel het type van de gebruikte fraseologische eenheden beïnvloedt (Biber et al., 2004), alsook de mogelijkheid om aandacht te besteden aan de linguïstieke output. Gedacht wordt namelijk dat men over het algemeen een hoger complexiteitsniveau behaalt met schrijven dan met spreken (Kormos, 2014; Manchón, 2014; Skehan, 2014).

Alhoewel onderzoek naar T2 Frans heeft aangetoond dat men een hogere kwantiteit aan fraseologische eenheden gebruikt wanneer de taalvaardigheid toeneemt (bijv. Forsberg, 2010), zijn er nog geen studies gedaan naar de complexiteit (diversiteit en verfijning) over verschillende taalvaardigheidsniveaus in T2 Frans, alsook het verschil in complexiteit tussen geschreven en gesproken taal. Deze studie dicht de kloof door te onderzoeken in hoeverre fraseologische com-

plexiteit een indicator is voor taalvaardigheid en ontwikkeling in T2 Frans (onderzoeksvraag 1), de mate waarin fraseologische complexiteitsmaten zich verhouden tot andere aspecten van taalkundige complexiteit (syntactische, lexicale en morfologische complexiteit) in T2 Frans (onderzoeksvraag 2) en in hoeverre complexiteit verschilt tussen geschreven en gesproken T2 producties (onderzoeksvraag 3).

Deze vragen worden beantwoord in een reeks empirische onderzoeken waarin zowel cross-sectionele als longitudinale corpora van gekoppelde geschreven en gesproken taken van dezelfde leerlingen worden onderzocht. Uit het onderzoek volgt dat zowel fraseologische als non-fraseologische complexiteitsmaten significante voorspellers zijn voor het taalvaardigheidsniveau (bepaald door experts middels holistische taalvaardigheidsbeoordeling) en dat fraseologische complexiteit geleidelijk verandert over tijd. In beide gevallen zijn deze effecten klein en lijken gemodereerd te zijn door communicatieve functie en de beperkingen van real-time productie in gesproken taal. Deze bevindingen laten het belang van een meer 'organische' benadering van complexiteit zien (Norris & Ortega, 2009), welke ook de effecten van aspecten als doeltaal of modaliteit in acht neemt wanneer de betekenis van complexiteitsmaten worden geïnterpreteerd.

Acknowledgements

Words cannot express my gratitude to my amazing supervisors, Magali and Alex. You have been a constant source of inspiration and I have grown enormously as a researcher thanks to your support and encouragement. Magali, thank you for your *very* detailed and rigorous feedback, for always pushing me to refine my ideas and for instilling an appreciation for 'real' Belgian chocolate (although I have to admit that I still love *Tony Chocolonely*). Alex, thank you for the thought-provoking discussions, for encouraging me to take risks, and for the lovely dinners at *Le Quartier Latin*.

I am also incredibly fortunate to have an examination committee composed of researchers whose work has been fundamental to this project. There's something quite special about getting to meet the people behind the articles and chapters that you cite on a regular basis and realizing that not only are they extremely brilliant researchers, but also incredibly kind people. Thank you very much to Thomas François who happily agreed to share his corpus data with a still wet-behind-the-ears Masters student way back before I knew I would ever end up at the UCLouvain. You constantly kept me on my toes by challenging me to maintain a high degree of methodological and statistical rigour. To Bastien De Clercq, thank you for your endless amounts of patience and guidance during panicked video calls and WhatsApp chats these past few years. I can always count on you for the perfect balance of empathetic, yet critical feedback. To Fanny Forsberg Lundell, whose work first ignited my interest in L2 French phraseology, it has been a real gift to be able to learn from such an inspiring researcher as yourself. I treasure the time I spent in

Stockholm as one of the highlights of the past four years. Marije Michel, while our paths didn't quite cross while I was still working in Groningen, I am so glad that we have been able to meet now. The time we had coffee at Black & Bloom was a real turning point in my thinking about many of the central ideas of this thesis. Lastly, to Doug Biber, it goes to show just how foundational your work has been that your name appears 143 times in the pages that follow (yes, I did a corpus search of my own thesis). Your voice has often been in the back of my mind while writing and I feel very fortunate to have the chance to discuss these ideas with someone I admire very much.

As a corpus-based researcher, I am indebted to those who have compiled the data upon which the studies in this thesis are based. Access to the *Leerdercorpus Frans* was generously provided by partners at universities in Belgium including Piet Desmet, Hans Paulussen and Frederik Cornillie at KULeuven; Gudrun Vanderbauwhede at the Université de Mons; as well as Annemie Demol and Pascale Hadermann at Universiteit Ghent. The LANGSNAP corpus was compiled at the University of Southampton by Rosamond Mitchell, Laurence Richard, Patricia Romero de Mills, Nicole Tracy-Ventura and Kevin McManus who have kindly made the data publicly available for research purposes. I would also like to thank our partners at the *Chambre de commerce et d'industrie de région Paris-Île-de-France* for providing access to data from the TEF exam. In particular, I would like to extend my heartfelt gratitude to Dominque Casanova for his help preparing and transferring the data and for his unending patience in responding to my many questions and requests for information.

Throughout the past four years, there have been many who have provided technical, theoretical and practical support and while this is far from an exhaustive list, I would like to express my appreciation specifically to colleagues at the CENTAL including Patrick Watrin and Hubert Naets as well as Alain Guillet at the SMCS. I also owe a huge debt of gratitude to my wonderful student assistants, Anthonya Delfosse, Héloïse Copin and Thibaut Mareschal for their hard work transcribing and annotating seemingly endless amounts of data. I look back with fondness at the many afternoons spent laughing over silly example sentences with you. I am also extremely grateful to those researchers

who have allowed me to pick their brains over the years including Bram Bulté, Stefan Gries, Jonas Granfeldt and Norbert Schmitt.

I could not have made it through the past few years without the support of colleagues at the CECL and the CLIN. To my doctoral "sibling," Rachel, I am so happy to have been able to share this journey with you. I have learned so much from our (often quite heated) discussions about complexity and I treasure the many ridiculous moments and adventures we have had trying to make it through this crazy thing called a PhD. Surviving the past four years was also made possible through a combination of donuts, coxinha, Basic Fit workouts, discussions of RuPaul's Drag Race and (virtual) coffee breaks shared with: Bert, Eva, Georgia, Gillian, Ily, Iris, Laura, Laurie, Pauline, Rory, Tanguy, Tove, Valérie, Yating and many others. Thank you also to Carles, Klara, Matti and Anna for making me feel welcome during my research stay at Stockholm University and for introducing me to the amazing concept of fika!

Completing a PhD also has its share of frustrating moments and I am so appreciative of my friends and family for always being there and for believing in me (even when I didn't believe in myself). To Stefan, our walks and runs in Noorderplantsoen were a constant lifesaver during the bleak moments of Covid-19. Amaranta, Paolo and Sally, thank you for always being a (virtual) shoulder to cry on and for being the best "chosen family" I could ask for. Finally, to my partner, Jeroen, thank you for putting up with my grumpy moments. Your love and patience have been a rock and you are always my biggest cheerleader. I'm so much looking forward to our post-PhD life together and to catching up on everything we've had to put on hold the past few years (there's a beach towel in Dubrovnik with our names on it!).

Lastly, I would be remiss if I didn't mention those who inspired me to pursue research in the first place. To my parents, Brian and Heather, thank you for encouraging me to be creative and curious about the world from a young age. To my wonderful high school French teacher, Alex Skene, thank you for igniting my passion for linguistics and for showing me how to be a kind and compassionate teacher. Last, but not least, thank you to my friends and former colleagues at the Uni-

versity of Groningen. Little did I know that a simple Google search for an MA program in Applied Linguistics in the midst of a cold, dark Japanese winter would lead to the beginning of such an exciting academic adventure. To my academic role models, Merel Keijzer, Remco Knooihuizen and Marjolijn Verspoor, thank you for first planting the idea of a PhD in my head and for convincing me that I could really do this!

Supplementary Materials

All supplementary materials for this thesis are hosted on an OSF repository (https://osf.io/w6xhk/). These include transcription guidelines, additional methodological details (e.g., topic-modelling procedures) and R scripts detailing the full statistical analyses that were conducted (e.g., model-fitting procedures). The OSF page also includes a link to the GitHub repository where the *fsca* package is hosted.

Abbreviations

C-Unit	Conversational Unit
CAF	Complexity Accuracy Fluency
CEFR	Common European Framework of Reference
CH	Cognition Hypothesis
FFD	French Frequency Dictionary
FS	Formulaic Sequence
LAC	Limited Attentional Capacity model
LANGSNAP	Languages and Social Networks Abroad Project
LCF	Leerdercorpus Frans
LT	Lexique Transdisciplinaire
MD	Multidimensional Analysis
PMI	Pointwise Mutual Information
POS	Part of Speech
RTTR	Root type-token ratio
SLA	Second Language Acquisition
T-Unit	Minimally Terminal Unit
TBLT	Task-based Language Teaching
TEF	Test d'évaluation de français
TTR	Type-token ratio

Chapter 1

Introduction

Introduction

The construct of L2 complexity has its roots in early attempts to define objective measures of development and proficiency (Hakuta, 1975; Larsen-Freeman, 1978; Larsen-Freeman & Strom, 1977; Nihalani, 1981). Following in the path of L1 acquisition studies (e.g., Hunt, 1965), L2 researchers measured lexical and grammatical aspects of learner productions in the hopes of finding a "developmental yardstick against which global (i.e., not skill nor item specific) second language proficiency could be gauged" (Larsen-Freeman, 1983: 287). Such measures included, for example, the diversity of vocabulary used or the length and degree of embedding of syntactic units. In contemporary second language research, these lexical- and syntactic-based indices became known as *complexity* measures because they were believed to relate to "the stage and elaboration of the interlanguage system" (Skehan, 1996: 46) and in particular to the size of the system (i.e., the number of syntactic and lexical elements) and the degree of organization within it (Skehan, 1998: 85). In spite of the many problems regarding the way it has been defined and operationalized (Biber et al., 2020, 2021; Bulté & Housen, 2012; Norris & Ortega, 2009; Pallotti, 2015), complexity is considered a key construct in the evaluation of L2 production, proficiency and development as part of the tripartite *Complexity Accuracy Fluency* (CAF) framework (Housen et al., 2012; Housen & Kuiken, 2009; Skehan, 1996, 2009b).[1] Within the CAF framework, the complexity of different linguistic subdomains (i.e., lexis, syntax, morphology) is often considered separately, which means that complexity at the *interface*

[1]Skehan (2009b) also includes lexis as a separate category (CALF) but for ease of reference and for consistency with the way that complexity is defined in Chapter 2, this thesis will refer to the framework simply as CAF.

of these domains is not taken into account.

Corpus linguistic evidence shows, however, that lexis and grammar are in fact quite inter-related (Goldberg, 2006; Römer, 2009; Sinclair, 1991; Stefanowitsch & Gries, 2003). This is exemplified by various multi-word, or *phraseological* phenomena including collocations (e.g., *strong + tea*), collostructions (e.g., *give* + DITRANSITIVE) and lexical bundles (e.g., *I don't think*) (Granger & Paquot, 2008; Gries, 2008). For L2 learners, mastery of co-occurrence and recurrence phenomena such as these is especially difficult (Paquot & Granger, 2012). For example, compared to native speakers and more advanced learners, beginners have been shown to under-use highly exclusive collocations (e.g., *densely + populated*) (Durrant & Schmitt, 2009; Granger & Bestgen, 2014) and have also been shown to exhibit less variety in the phraseological units they use (Erman et al., 2015; Forsberg, 2010), relying instead on a smaller set of phraseological "teddy bears" (Nesselhauf, 2005).

In an attempt to unite L2 complexity research with L2 phraseology research, Paquot (2019) proposed the construct of *phraseological complexity*, which she defined as the diversity and sophistication of phraseological units used in learner production. In the case of L2 English, these two dimensions have been shown to be associated with written proficiency (Garner et al., 2018, 2019; Garner, 2020; Granger & Bestgen, 2014; Jiang et al., 2021; Kim et al., 2018; Paquot, 2018, 2019; Zhang & Li, 2021) and to a limited extent also oral proficiency (Eguchi & Kyle, 2020; Kyle & Crossley, 2015; Uchihara et al., 2021; Zhang et al., 2021). At the same time, there is evidence to suggest that there may be differences in the phraseological complexity of speech as compared to writing. For example Paquot et al. (in press) as well as Tavakoli and Uchihara (2020) found that the use of highly associated collocations (a measure of phraseological sophistication) was actually negatively correlated with oral proficiency. This body of research therefore suggests that phraseological complexity measures may indeed be a useful 'yardstick' for written proficiency and development, at least in the case of L2 English. However, the applicability of the construct of phraseological complexity to languages other than English has seldom been empirically tested (cf. Edmonds & Gudmestad, 2021;

Rubin et al., 2021; Siyanova-Chanturia & Spina, 2020), nor has any study directly compared the phraseological complexity of oral and written production.

The first aim of this thesis is therefore to address this gap by investigating the extent to which phraseological complexity measures are predictive of proficiency and development in L2 French. As suggested by Stengers et al. (2011), in more synthetic languages than English, morphology may interact with phraseology because learners need to know a large number of different inflected variants for each word combination. This means, for instance, that despite being aware of a collocation between two lemmas, they may not be able produce the required grammatical inflection. As a result, their text may demonstrate high levels of phraseological complexity but low morphosyntactic accuracy, which may negatively affect a rater's perception of the learner's proficiency level.

The second aim of this thesis is to investigate how phraseological complexity compares to other types of linguistic complexity (i.e., lexical, syntactic, morphological) as an index of development and proficiency in L2 French. In the case of L2 English, Paquot (2019), showed that measures of phraseological complexity were stronger predictors of advanced proficiency than measures of lexical or syntactic complexity. However, given the inflectional properties of French, it may be the case that other types of complexity (e.g., morphological complexity) continue to be predictive of advanced proficiency in French.

The third aim of this thesis is to investigate the extent to which phraseological complexity differs in written versus spoken L2 production. As noted above, compared to the number of studies that have looked at L2 writing, there is relatively little research on phraseological complexity in L2 speech and no study has directly compared phraseological complexity across modes. This is important because writing and speaking differ in a number of important ways, but especially in terms of how much attention learners can pay to output (Kormos, 2014; Skehan, 2014) and in terms of which linguistic features tend to be used (Biber & Conrad, 2019), both of which may impact phraseological complexity.

The research questions for this thesis are thus:

> **RQ1.** To what extent is phraseological complexity an indicator of proficiency and development in L2 French?
> **RQ2.** To what extent does phraseological complexity relate to other aspects of linguistic complexity as tapped by the current repertoire of measures of syntactic, lexical and morphological complexity in L2 French?
> **RQ3.** To what extent does phraseological complexity differ in oral versus written L2 production?

The above research questions are addressed through a series of empirical, corpus-based studies, which make up the main chapters of this thesis. Chapter 2 provides the theoretical framing for these articles, placing them in the context of previous research that has been done on L2 complexity, L2 phraseology and the influence of modality on complexity.

Before attempting to compare phraseological complexity across modes, it was important to first determine whether previous findings for L2 English would extend to the context of L2 French by carrying out a partial replication study of Paquot (2018, 2019). What quickly became apparent however, is that, compared to the wide range of tools available to calculate complexity measures for L2 English (e.g., Kyle & Crossley, 2015; Lu, 2010), the availability of such tools for L2 French was much more limited. This meant that in order to be as faithful as possible to the Paquot's original studies, it was necessary to develop, test and implement such a tool from scratch. The first empirical study in this thesis (Chapter 3) reports on this process. While it does not directly contribute to answering the main research questions in and of itself, the development of the tool made it possible to compare phraseological complexity to syntactic complexity in the subsequent chapter.

The second empirical study (Chapter 4) reports on the partial replication of Paquot (2018, 2019). The principle aim of this study is to determine the extent to which phraseological complexity measures, originally developed for L2 English, are also applicable in the context of L2 French. The data for this study come from the *Leerdercorpus Frans*

(Demol & Hadermann, 2008; Vanderbauwhede, 2012), a corpus of argumentative essays written by university-level Dutch learners of French. Each text in the corpus was independently assessed by trained raters and assigned a holistic proficiency score. On the basis of this corpus, several phraseological, syntactic, lexical and morphological complexity measures were calculated in order to determine the relative contribution of each type of complexity to the holistic evaluations of written proficiency.

The third study (Chapter 5) investigates how phraseological complexity develops over time in written versus oral production. The data for this longitudinal study come from the Languages and Social Networks Abroad (LANGSNAP) project (Tracy-Ventura et al., 2016), which followed 29 university-level learners of French before, during and after a sojourn abroad in France. The participants completed three production tasks at multiple time points over the course of 21-months: an argumentative essay, an oral narrative and an oral interview. This study compares the development of phraseological complexity over time in the three tasks.

The fourth and final study (Chapter 6) takes a cross-sectional approach to phraseological complexity, using data from the *Test d'évaluation de français* (Chambre de commerce et d'industrie de Paris, 2010). This study aims to determine the extent to which the phraseological and lexical complexity of oral versus written exam productions contribute to the scores given to those productions by professional raters. This study also takes a more applied perspective and discusses the contribution that phraseological complexity research can make to the field of language assessment.

Chapter 7 summarizes the results of the three empirical studies and attempts to answer the research questions listed above. This chapter also reviews some of the methodological limitations of the thesis and provides suggestions for future research.

Chapter 2

State of the art

Overview

The current chapter lays out the theoretical foundation for the empirical studies that form the core of this thesis. First, Section 2.1 provides a general overview of the construct of L2 complexity and reviews how complexity measures have typically been used in L2 French research. Section 2.2 then introduces Paquot's (2019) construct of *phraseological complexity* and reviews the current body of research that has been done on phraseological complexity in L2 English, drawing parallels with studies of phraseology that have been done in L2 French. Finally, Section 2.3 turns to the issue of modality and shows how the relationship between complexity and modality has been seen through two lenses: as a task related factor (Section 2.3.2) and as an aspect of register variation (Section 2.3.3). Section 2.4 summarizes this state of the art and highlights how the empirical studies in this thesis attempt to fill an important gap in the phraseological complexity literature.

2.1 The construct of L2 complexity

As Bulté and Housen (2012) point out, there are two common uses of the term *complexity* in applied linguistics. On the one hand, complexity can refer to linguistic features that require more cognitive processing (Hulstijn & de Graaff, 1994) or that are more difficult for L2 learners to acquire (Szmrecsanyi & Kortman, 2009). Bulté and Housen (2012) refer to this as *relative complexity* due to the fact that such factors are subjective and may differ across learners. The other type of complexity is *absolute complexity*, which is defined in terms of objective properties of a linguistic feature or the linguistic system more broadly (Miestamo, 2008). Definitions of absolute complexity in applied linguistics often refer to the philosopher Rescher (1998: 1), who describes the complexity of a system as "the quantity and variety of its constituent elements and of the interrelational elaborateness of their organizational and operational make-up." Along these lines, Bulté and Housen (2012: 24) define absolute complexity in linguistics as the "number of discrete components that a language feature or a language system consists of, and as the number of connections between the different components."

Clearly distinguishing between absolute and relative forms of complexity is important in order to avoid defining one form of complexity in terms of the other (Housen et al., 2019; Pallotti, 2015), the danger being that doing so can lead to circular argumentation (e.g., calling a structure complex because it is acquired later but at the same time claiming that it is acquired later because it is complex). That being said, it is important to note that relative and absolute complexity do often intersect in practice (De Clercq, 2016). As Rescher (1998: 8) notes, "the more complex something is, the more difficulty we have in coming to grips with it and the greater the effort that must be expanded for its cognitive and/or manipulative control and management." This means that linguistic features or systems that are complex in an absolute sense *may* also be more cognitively difficult for learners or may be acquired later in development. It is important to keep in mind though that this is not a deterministic relationship (Pallotti, 2009, 2015). Sometimes complexity in an absolute sense may actually lead to less cog-

nitive difficulty. To illustrate this, Pallotti (2015) uses the example of a Sudoku puzzle with 25 digits compared to one with only 18. Although the Sudoku puzzle with more digits is more complex in an absolute sense (i.e., having more constituent components), it is less cognitively challenging to solve than the puzzle with fewer digits. Likewise, when it comes to the complexity of linguistic features, just because something is more complex in an absolute sense, it does not *automatically* follow that it is more cognitively challenging or difficult for learners.

When discussing L2 complexity, Bulté and Housen (2012) also argue for the importance of clearly articulating the difference between three different levels of analysis (as shown in Figures 2.1 and 2.2):

1. At a *theoretical* level, complexity can be seen either as an abstract property of linguistic systems, which they refer to as *system complexity* (e.g., the number of words, syntactic structures, morphological inflections known to a learner) or as the complexity of individual linguistic features, which they refer to as *structure complexity* (e.g., the degree embeddedness of a syntactic structure).
2. At an *observational* level, it can be seen as the concrete characteristics of learner productions in which such properties are manifest (e.g., constructs of density, diversity, sophistication, compositionality etc.).
3. Finally, at an *operational* level, complexity can be seen through the specific measures used to tap into such characteristics (e.g., type-token ratios or the ratio of subordinate clauses to matrix clauses).

Just as with the distinction between relative and absolute complexity however, these three levels of analysis often intersect in practice because depending on how complexity is operationalized, measures can tap into more than one observational or theoretical construct, as discussed in the following section.

2.1.1 Operationalizing L2 complexity

To date, most measures of L2 complexity have focused on lexical, syntactic and morphological complexity (Bulté & Housen, 2012; Wolfe-

Quintero et al., 1998). This section provides a brief overview of how complexity has typically been operationalized within each of these domains.

Lexical complexity. As shown in Figure 2.1, at the observational level, lexical complexity is typically represented by the constructs of density, diversity, compositionality and sophistication (Bulté et al., 2008; Bulté & Housen, 2012; Skehan, 2003) and measured through ratios of lexical words to function words, the ratio of unique words (types) to the total number of words used (tokens) (i.e., the type-token ratio (TTR), Templin, 1957),[1] ratios of morphemes or syllables to words or proportions of low-frequency words (frequency being determined on the basis of an L1 reference corpus) (Bulté et al., 2008; Wolfe-Quintero et al., 1998).

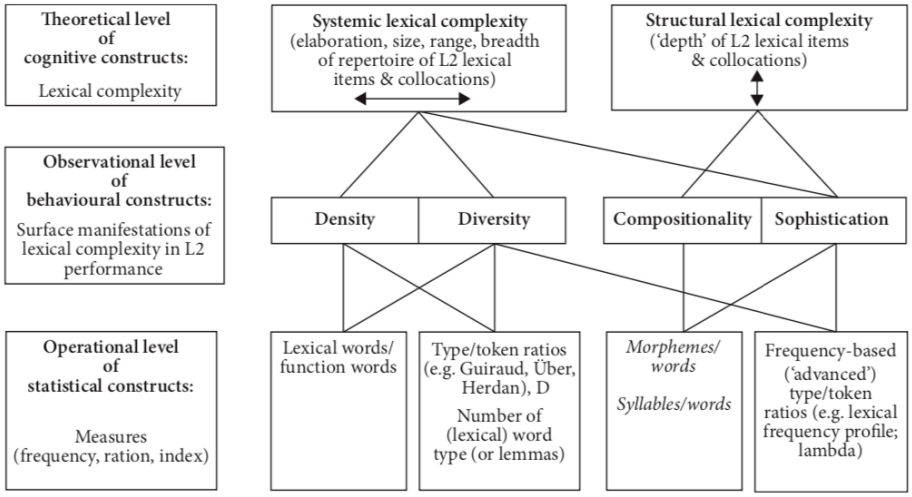


Figure 2.1: Lexical complexity at different levels of construct specification (Bulté and Housen, 2012: 28)

As shown by the various intersecting lines in Figure 2.1, these measures often tap into more than one construct. TTR-based measures, for example, can tap into both density and diversity. Likewise, frequency-

[1] Because raw TTR is often strongly correlated with text length, various solutions have been proposed to compensate for this including (but not limited to) Johnson's (1944) mean segmental type-token ratio (MSTTR), Guiraud's (1954) root type-token ratio (RTTR), Carroll's (1964) corrected type-token ratio (CTTR), Covington and McFall's (2010) moving average type-token ratio (MATTR), McCarthy and Jarvis's (2007) Voc-D/D and McCarthy and Jarvis' (2010) Measure of Textual Lexical Diversity (MTLD), which have each been used to a varying extent in L2 complexity research.

based measures can reflect both diversity and sophistication. In addition, some of these constructs overlap with both relative as well as absolute forms of complexity. As argued by Pallotti (2015), both density and sophistication represent aspects of relative complexity. His claim is that there is nothing inherently complex (in an absolute sense) about the use of content words or relatively rarer words such as 'tar' as opposed to function words (e.g., 'the') or more frequent words (e.g., 'car'). Moreover, the use of any given word is at least partially determined by individual factors (e.g., the amount of previous exposure a learner has received to the word). At the same time, there is also an argument to be made for sophistication as an aspect of absolute complexity. As described by Bulté et al. (2008: 279), sophistication implies "knowledge of semantic relations (e.g., hyponymy, hypernymy, synonymy, antonymy) and, hence, of different but related lexical alternatives for referring to a referent." When a learner uses an infrequent word, this may therefore indicate (1) that he or she has selected this word from a broad repertoire of words and (2) that the choice of the word was made by considering the semantic relations between different words. Framed in this way, the use of infrequent words could be said to be indicative of both system complexity (i.e., more different words) and structure complexity (i.e., the dense network of relations between words), aspects of absolute complexity. This is shown in Figure 2.1 by the lines connecting the observational construct of sophistication with theoretical constructs of both system and structural complexity. What is clear from these examples is that the links between lexical complexity measures and observational/theoretical constructs are not always straightforward.

Syntactic and morphological complexity. As shown in Figure 2.2, there is also overlap in the way that observational and theoretical constructs of syntactic and morphological complexity are represented by different types of measures. At the most basic level, syntactic and morphological complexity can be divided into the constructs of diversity and sophistication. The most commonly used measures of these constructs include the length of structures (e.g., mean length of clause, T-unit, noun phrase), proportions of one type of syntactic or morphological structure to another (e.g., ratio of dependent/subordinate clauses

to matrix clauses) and diversity indices such as the *Syntactic Diversity Index* (SDI, De Clercq & Housen, 2017) or the *Morphological Complexity Index* (MCI, Brezina & Pallotti, 2019) (not pictured).

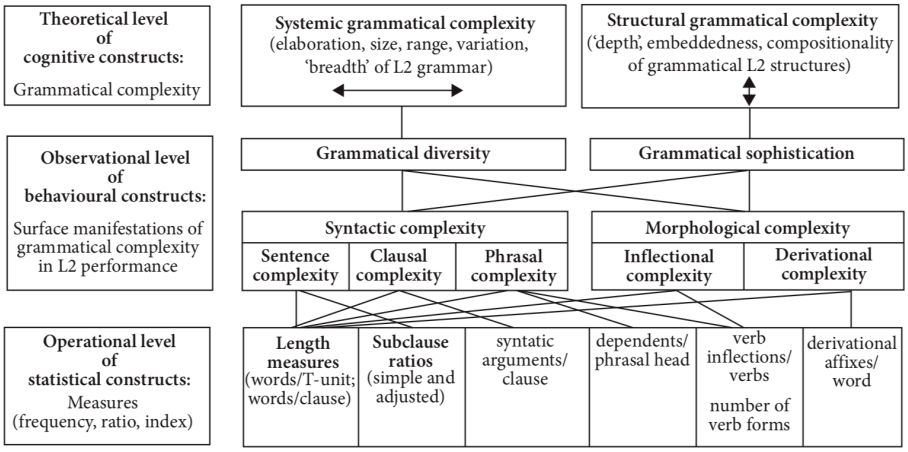


Figure 2.2: Grammatical complexity at different levels of construct specification (Bulté and Housen, 2012: 27)

It is important to point out here that the use of these measures as indices of 'syntactic' complexity is not uncontroversial. For one, they have been criticized for prioritizing clause-level complexity at the expense of other types of grammatical phenomena such as non-finite clauses and non-clausal phrases (Biber et al., 2020, 2021; Norris & Ortega, 2009). In addition, length-based measures in particular are problematic because they collapse many different structural phenomena together. As explained by Norris and Ortega (2009: 561), "the addition of subordinate clauses can lengthen an utterance or T-unit, but so can adding adjectives and prepositional phrases that pre- or postmodify nouns, or adding non-finite verb phrases that modify other elements via nonsubordinating clausal means." Length-based measures therefore reflect complexity at several hierarchical levels (i.e., sentence, clause, phrase), as shown in Figure 2.2. Perhaps most problematically, these so-called omnibus measures disregard relevant differences between grammatical structure (e.g., dependent clauses) and syntactic function (e.g., finite adverbial clauses, relative clauses) (Biber et al., 2020, 2021), which comes at the expense of "linguistically inter-

pretable descriptions" of a learner's performance or proficiency. Such differences are highly relevant when comparing the types of grammatical features used at different proficiency levels (Biber et al., 2011) or across different registers (e.g., Biber et al., 1999). In particular, "structural characteristics and syntactic functions are equally important for describing the complexity differences among spoken and written registers" (Biber et al., 2021: 461), which is an issue that will be revisited in Section 2.3.3.

As Bulté and Housen (2012) point out, another difficulty with syntactic complexity measures is that they often tap into both system complexity and structure complexity. The use of longer syntactic constituents, for example, can be indicative of a learner with a larger repertoire of syntactic structures while at the same time, the constituent itself may be longer because of embedded sub-constituents, reflecting structural complexity. Again, this is shown in Figure 2.2 through the multiple intersecting lines connecting measures to observational and theoretical constructs. Moreover, measures of syntactic complexity also represent both absolute and relative forms of complexity. For instance, following the recommendation of Norris and Ortega (2009), most contemporary studies of L2 complexity (e.g., Lu, 2010) tend to target three different types of syntactic phenomena: coordination, subordination and phrasal elaboration. This is based on the theoretical foundation of systemic functional linguistics (Halliday & Martin, 1993; Halliday & Matthiessen, 1999), which claims that linguistic development proceeds largely through three stages. First, learners join words and propositions together through juxtaposition and coordination (i.e., parataxis), then they also learn to subordinate ideas to each other (i.e., hypotaxis) before eventually being able to incorporate grammatical metaphor (e.g., 'she is clever' becoming 'her cleverness'). From a purely absolute complexity point of view, subordination *is* structurally more complex than coordination because it requires hierarchical organization. However, both clausal and phrasal elaboration involve structural embedding (either of clauses or phrases) (see discussion in Biber et al., 2021: 470) so in an absolute sense, one is not necessarily more complex than the other; the major difference between them being developmental timing, an aspect of

relative complexity (Bulté & Housen, 2012; Pallotti, 2015).

The brief discussion here is meant to illustrate that distinctions between relative or absolute forms of complexity, or between observational and theoretical constructs may be helpful in principle (e.g., to avoid circular reasoning) but in practice, the way that complexity measures are operationalized means that it is not always possible to maintain these strict divisions. Likewise, there are important questions regarding the validity of many commonly-used complexity measures. These issues continue to be the source of ongoing debate within L2 complexity research (e.g., Biber et al., 2020, 2021; Bulté & Housen, 2012; Housen et al., 2019; Norris & Ortega, 2009; Ortega, 2003, 2012; Pallotti, 2015) and it is an issue that cannot be easily resolved here, especially in relation to phraseological complexity (as discussed in Section 2.2). That being said, despite potential inconsistencies regarding the links between complexity measures and the constructs believed to underlie them, these measures have often proved to be useful indices of L2 performance and proficiency, which is the subject of the following section.

2.1.2 Complexity measures in L2 French

When it comes to how complexity measures have been used in L2 research, Ortega (2012) distinguishes between three broad purposes: as indices of development, as indices of proficiency and as a way to describe performance based on task-related factors. The following section briefly outlines how complexity measures have been used for these purposes in L2 French.

Indices of development. In terms of longitudinal development, Bulté et al. (2008) observed that over a period of three years, the oral narratives produced by high school-level learners of French, significantly increased in both lexical diversity and sophistication. Both Lissón and Ballier (2018) as well as Macaro and Masterman (2006), likewise report significant effects of time on the lexical diversity of written learner productions (for absolute beginners and intermediate-level learners respectively). Another example of the use of complexity as a developmental index in L2 French comes from the LANGSNAP project (Tracy-

Ventura et al., 2016), which collected oral and written production data from 29 university-level learners of French before, during and after a 9-month sojourn abroad in France (see overview in Section 5.2.1). Over the study period, there were significant increases in lexical diversity of oral interviews (Mitchell et al., 2017) and oral narratives (McManus et al., 2021). In addition, there were small but significant increases in the syntactic complexity (clauses per T-unit, mean length of T-unit) in both oral narratives (McManus et al., 2021) and argumentative essays (Mitchell et al., 2017).[2] Finally, Kern and Schultz (1992) report a significant increase in syntactic complexity (mean length of T-unit) in the writing of university-level learners of French over the course of one year of instruction.

Indices of proficiency. Cross-sectional research has also used complexity measures to investigate differences between groups in L2 French. Harley and King (1989), for example, compared the lexical complexity of L1 and L2 written productions (high school-level French immersion students and L1 peers) and found that the learner texts exhibited significantly less lexical diversity and sophistication as compared to the native texts. Comparing across proficiency levels of L2 learners, Gyllstad et al. (2014) found that emails and stories written by high school-level learners that were rated at the B levels of the Common European Framework of Reference (CEFR, Council of Europe, 2001) had significantly higher levels of syntactic complexity (mean length of T-unit, mean length of clause, clauses per T-unit) than texts at the A level. Similar results have also been reported for oral production. David (2008), for example, compared the lexical diversity of oral interviews conducted with high school learners of French in the United Kingdom (the French Learner Language Oral Corpus, Myles & Mitchell, 2011) and found that except for the final two years, there was a significant increase in lexical diversity with each year of study. Along the same lines, Welcomme (2013) analyzed a corpus of oral narratives collected from high school-level learners in Belgium and found significant differences between the learners in different years of study in the proportion of juxtaposed clauses, coordinated clauses and subordinated clauses (both finite and non-

[2] Not all complexity measures were calculated for each task type.

finite). In a series of studies based on a larger sample from the same corpus, significant differences between the learner groups[3] were also reported for various measures of lexical diversity and sophistication (De Clercq, 2015), syntactic diversity and elaboration[4] (De Clercq & Housen, 2017) and morphological diversity (De Clercq & Housen, 2019). Complexity measures have also been found to distinguish between more advanced groups of learners. Comparing first- and final-year university students in the United Kingdom, Tidball and Treffers-Daller (2007, 2008), for instance, found significant differences between the groups in terms of both lexical diversity and sophistication (see also Treffers-Daller, 2013). Likewise, Lindqvist et al. (2011) as well as Forsberg Lundell and Lindqvist (2014) found that interviews conducted with university-level learners and non-native speakers who were long term residents in France also differed with respect to lexical sophistication. Finally, similar differences are reported by Ovtcharov, Cobb and Halter (2006) for adult civil servants rated as intermediate and advanced on the *Évaluation de langue seconde* (ELS) exam.

Differences in task conditions. Complexity measures have also been used as a way to test differences in task conditions in L2 French. One example is that of Kuiken and Vedder (2012) who manipulated the number of elements that learners needed to take into account in letters to a friend about a holiday destination. The results showed that the more cognitively *complex* condition (i.e., the one with more elements to consider) led to higher levels of lexical sophistication than the less complex task. The authors also measured syntactic complexity (clauses per T-unit, dependent clauses per T-unit) and lexical diversity (CTTR) but neither of these measures were found to differ between tasks. Another study to investigate the effect of task conditions on complexity was that of Granfeldt (2007), who examined the effect of modality on complexity measures, finding higher levels of lexical diversity (D) in writing as compared to speaking but no significant difference for syntactic complexity (clauses per T-unit). As this study is directly relevant for the relationship between complexity and modal-

[3] In these studies proficiency groups were defined in terms of both year of study and by an additional criteria of morpho-syntactic accuracy.

[4] In this study elaboration refers to length-based measures and the ratio of clauses to AS-units.

ity, it is discussed further in Section 2.3.2.

The studies outlined above show that measures of lexical, syntactic and morphological complexity can be useful indices of development, proficiency and, to some extent, the effect of task characteristics in L2 French. This is demonstrated by the fact that such measures have been observed to increase with more time and exposure to French (Bulté et al., 2008; Kern & Schultz, 1992; Lissón & Ballier, 2018; Macaro & Masterman, 2006; McManus et al., 2021; Mitchell et al., 2017), distinguish between *a priori* determined proficiency levels (David, 2008; De Clercq, 2015; De Clercq & Housen, 2017, 2019; Forsberg Lundell & Lindqvist, 2014; Gyllstad et al., 2014; Ovtcharov et al., 2006; Tidball & Treffers-Daller, 2007, 2008; Welcomme, 2013) and differentiate between experimental task conditions (Granfeldt, 2007; Kuiken & Vedder, 2012). As Pallotti (2021) points out, it is important to be keep in mind that this does not, on the face of it, speak to the validity of such measures as indices of complexity but rather to the "practical usefulness" of such measures in tracking development or discriminating between groups.

At the same time, the complexity measures used in the above studies did not always prove useful for these purposes. For instance, in the series of studies conducted on the basis of the Belgian corpus of oral narratives, although significant differences were observed between some proficiency levels, none of the lexical, syntactic or morphological complexity measures were found to differ significantly between the two highest-level groups in those studies (De Clercq, 2015; De Clercq & Housen, 2017, 2019). Similarly, Lindqvist et al. (2011) as well as Forsberg Lundell and Lindqvist (2014) found that although there were differences in lexical sophistication between the lowest- and highest-level learners in their corpus, no such significant differences were found between the highest level learners (long-term residents in France) and native speakers. This suggests that the lexical, syntactic and morphological complexity measures employed in those studies were not able to capture differences between learners at the upper proficiency levels. Put another way, measures that target the complexity of individual words, syntactic structures or morphological inflections seem to be missing out on an important aspect of the complexity of the inter-

language system. Specifically, what is missing from these measures is the complexity of phenomena at the interface of these domains. The following section will show how Paquot's (2019) construct of *phraseological complexity* aims to fill this gap by measuring complexity at the interface of lexis and grammar.

2.2 Phraseological complexity

To illustrate why 'traditional' measures of complexity may not be sufficient on their own, consider the sentences in Example 2.1:

Example 2.1. Sentences with the high-frequency verb *meet* (Paquot, 2019: 122)

a) I'll **meet** you in the bar tomorrow.
b) I **met** up with John as I left the building.
c) This app has different versions to **meet** different needs.
d) To **meet** expectations, several initiatives have been taken.
e) If you **meet** your target, congratulate yourself.
f) 'Here I believe my brother has **met** his Waterloo,' she murmured.
g) There is more than **meets** the eye.
h) Many students are finding it difficult to make ends **meet**.
i) Nice to **meet** you!
j) It's a pleasure to **meet** you.

As Paquot (2019) points out, because measures of lexical complexity are normally based on single words (or lemmas/lexemes), they are not able to fully capture the complexity of the above example. The repetition of the same verb *meet* in every sentence would contribute to low levels of lexical diversity and because it is a high-frequency verb, it does not contribute to high levels of lexical sophistication based on the proportion of low-frequency words. Similarly, the most commonly used measures of syntactic complexity cannot account for the variety of different structures that the verb appears in, which include verb + direct object (e.g., *meet + needs*, *meet + expectations* etc.), verb + particle (e.g., *meet + up*), non-finite verbal complement (e.g., *to meet*...) and a finite dependent clause (e.g., *if you meet*...). This is not meant to imply that lexical or syntactic measures are somehow deficient. Rather,

this example illustrates that there is a need to expand the scope of L2 complexity measures in order to account for the complexity of phenomena at the interface between traditional linguistic domains such as lexis and grammar (for a similar argument see Housen et al., 2019). This is because the strict separation of lexis and grammar in L2 complexity research does not reflect the reality of language use (Römer, 2009).

Corpus linguistic evidence has shown that certain words and grammatical structures appear together much more often than would be "expected on the basis of chance" (Gries, 2008: 6) or are repeated together at high frequencies (Biber et al., 1999). This means that words do not fill grammatical structures randomly, but that there are meaningful patterns to the ways that words combine into phrases. This is perhaps best captured by Sinclair's (1991: 110) idiom principle: "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments." Often cited along these lines is Halliday's (1966) classic example of *strong tea* rather than *powerful tea*. While both are grammatically correct and represent basically the same meaning, the latter is unidiomatic for most native speakers and rarely if ever attested in corpora.

Although it is difficult to get a precise estimate of the extent to which language is made up of such idiomatic expressions due to the various ways that these phenomena have been categorized and operationalized (see for example Granger & Paquot, 2008; Wray, 2000), Erman and Warren (2000) suggest that up to half of language may be constrained by Sinclair's (1991) idiom principle. Going even farther, proponents of usage-based approaches to second language acquisition might argue that, to a certain extent, all language is based on statistical associations (N. C. Ellis & Wulff, 2015). This can be at the level of individual words (e.g., the semantic association of the word 'squirrel' with the meaning of a bushy tailed rodent), between multiple words (e.g., collocations like *strong + tea*), between words and grammatical structures (e.g., the association of 'give' with the DITRANSITIVE, Stefanowitsch & Gries, 2003) or even, as proposed by construction grammarians (e.g., Goldberg, 2006), between lexically unspecified argument structures

and semantic meanings (e.g., the DITRANSITIVE being associated with the meaning of 'transfer').

To account for complexity at the lexis-grammar interface, Paquot (2019: 124) proposed the construct of *phraseological complexity*, which she defined as the diversity[5] and sophistication of phraseological units in learner production (following Ortega, 2003). The term *phraseological unit* refers to "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: 6). This definition is purposefully broad to include the co-occurrence of words *within* specific syntactic structures (e.g., the direct object relation *meet + needs* or the adjectival modifier relation *strong + tea*), the co-occurrence of words *and* grammatical structures (e.g., meet + TRANSITIVE) as well as patterns of recurrence such as lexical bundles (e.g., 'I don't think') but does not include lexically unspecified constructions (e.g., DITRANSITIVE + semantic meaning of transfer).

2.2.1 Operationalizing phraseological complexity

The two dimensions of phraseological complexity, diversity and sophistication, originate from the way that complexity has commonly been operationalized in L2 research (Bulté & Housen, 2012; Ortega, 2003). As with lexical or syntactic complexity, these are observational constructs in which system and structure complexity may be manifest (see Figure 2.3) and are used to describe aspects of underlying linguistic competence that may be indicative of proficiency and development. In what follows, these two dimensions are each briefly described in more detail.

Phraseological diversity is meant to reflect the size of a learner's underlying phraseological repertoire. In that way, it primarily taps into system complexity. Just as with lexical diversity, Paquot (2019) first

[5]The original definition uses the term *range* instead of diversity but I have used the word diversity here, following the conventional use of the term in relation to lexical complexity.

proposed that this could be operationalized using type-token ratios or transformations such as Guiraud's (1954) root type-token ratio (RTTR). Evidence for the potential usefulness of this construct for describing L2 proficiency and development comes from studies showing that learners tend to over-use a relatively limited set of collocational "teddy bears" (e.g., *have + time*, *have + problem*) relative to native speakers (e.g., Altenberg & Granger, 2001; Hasselgren, 1994; Kaszubski, 2000; Nesselhauf, 2005). Studies of L1 and L2 lexical bundle use have similarly found that although learners use more bundles overall, they use fewer types (i.e., unique bundles) compared native speakers (Ädel & Erman, 2012; Chen & Baker, 2010; De Cock, 2000). Taken together, this suggests that using a more diverse set of phraseological units may be indicative of L2 proficiency and development.

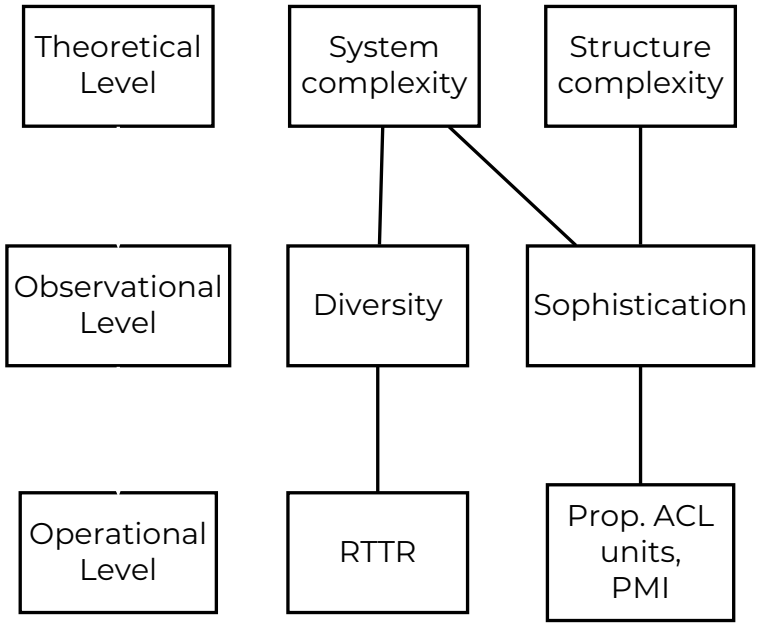


Figure 2.3: Phraseological complexity at different levels of construct specification and Paquot's (2019) proposed operationalizations

Phraseological sophistication represents a learner's use of "word combinations that are appropriate to the topic and style of writing rather than just general, everyday vocabulary" (Paquot, 2019: 125). This might include the use of technical terms, or combinations that

allow writers to more precisely express specific meanings (following Read, 2000). As with lexical sophistication, this can be said to reflect aspects of both system and structure complexity. On the one hand, the use of a sophisticated phraseological unit may be indicative of a broad repertoire of phraseological units in a learner's mental lexicon, and thus a reflection of system complexity. On the other hand, just as lexical sophistication represents knowledge of the semantic relations between words and related alternatives (Bulté et al., 2008), phraseological complexity represents the degree of interconnections between items in the mental lexicon (i.e., between words or between words and syntactic structures). In that sense, sophistication could also be said to represent structure complexity.

Paquot (2019) originally proposed two ways that phraseological sophistication could be operationalized. First, she suggested that it could be operationalized as the proportion of "academic phraseological units" (e.g., those from the Academic Collocation List, Ackermann & Chen, 2013). This would measure the ability of learners to draw on units that are more specific to an academic register rather than those typical of more common, everyday usage. The second way that Paquot recommended to operationalize sophistication was through the pointwise mutual information (PMI, Church & Hanks, 1990) of units calculated on the basis of an L1 reference corpus. PMI represents the degree of contingency between two items (i.e., whether two items are more frequently found together or separately). As explained by N.C. Ellis, Simpson-Vlach and Maynard (2008: 380), phraseological units with a high PMI are those "with much greater coherence than is expected by chance, and this coherence tends to correspond with distinctive function or meaning as well as grammatical well-formedness as a complete phrase." Examples of phraseological units with a high PMI (>3) from Paquot's (2019) study include: *overwhelming + majority*, *statistically + significant* and *arouse + curiosity*. In contrast, phraseological units with a low PMI (1) include: *final + part*, *also + possible*, *have + function*.

It is important to note that while there is an argument to be made that PMI represents absolute complexity (e.g., evidence of a broader repertoire of phraseological units, interconnections between words

or grammatical structures), one cannot ignore the fact that PMI also taps into aspects of relative complexity as well. Just as with frequency-based lexical complexity measures, it is not always easy to draw a strict division between absolute and relative complexity when it comes to PMI. This is because there is a relationship between PMI and how quickly phraseological units are processed. As shown by N. C. Ellis et al. (2008), units with a high PMI tend to be processed faster by native speakers on a variety of receptive and productive tasks (e.g., grammaticality judgement tasks, read aloud tasks). The processing of such units by learners, on the other hand, tends to be more related to the frequency of the units rather than PMI. As N. C. Ellis et al. (2008) point out, building up a mental representation that is highly attuned to co-occurrence phenomena requires an enormous amount of linguistic input, including a mental tabulation not just of how often a specific phraseological unit has been encountered (i.e., the frequency), but also how often each of the internal elements have been encountered elsewhere (i.e., contingency). For L2 learners, they may simply have not received enough input to form these contingency relationships in the mental lexicon. In part, this is precisely why PMI is a useful measure of proficiency and development: it reflects how aware a learner is of distributional patterns in the input as evidenced through their use of highly contingent phraseological units in production (Gries & Ellis, 2015).

In addition to the statistical properties of PMI (i.e., the fact that it represents highly exclusive phraseological units) and the psycholinguistic rationale as described above, Paquot's selection of PMI as a measure of phraseological sophistication was also based on previous research showing its usefulness as an index of L2 proficiency. For example, Durrant and Schmitt (2009) tested whether PMI could distinguish L1 and L2 written production. While this was certainly not the first study to compare the phraseology of L1 and L2 production (see for example De Cock et al., 1998; Granger, 1998; Howarth, 1998; Nesselhauf, 2005), this study was the first to measure PMI using an external L1 reference corpus. The authors manually extracted all directly adjacent premodifier-noun (i.e., adjective + noun and noun + noun) word pairs from a corpus of L1 and L2 writing and calculated the frequency and PMI of each item

in the British National Corpus (BNC). Consistent with the results of N. C. Ellis et al. (2008), Durrant and Schmitt (2009) found that the L1 writers tended to use more collocations with a higher PMI score whereas the L2 writers tended to use and repeat a smaller set of highly frequent collocations, which as the authors point out, suggests that "the learners are quick to pick up highly frequent collocations, but less common, strongly associated items … take longer to acquire" (p. 175). Using a similar methodology, Granger and Bestgen (2014) compared the collocations used in argumentative essays written by intermediate versus advanced learners (texts rated at the B and C levels of the CEFR). In line with expectations, the texts that were rated as more advanced had, in general, a higher proportion of high PMI collocations. The studies of Durrant and Schmitt (2009) and Granger and Bestgen (2014) therefore provided evidence that when operationalized as PMI, phraseological sophistication could be a useful index of L2 proficiency.

In two empirical studies, Paquot (2018, 2019) tested whether her two hypothesized dimensions of phraseological complexity, that is, diversity and sophistication, would be predictive of holistic evaluations (CEFR B2 to C2) of L2 research papers in Linguistics written by French learners of English. Using a method called dependency parsing, Paquot automatically extracted all adjectival modifiers (adjective + noun), adverbial modifiers (adverb + verb/adjective/adverb) and direct objects (verb + noun) from the learner corpus. Unlike the methods used by Durrant and Schmitt (2009) and Granger and Bestgen (2014), this method allows even long-distance relations to be extracted (i.e., those with many intervening words such as verb + direct object structures). Phraseological diversity was operationalized as the type-token ratio of the units with Guiraud's (1954) transformation applied (RTTR, Root Type Token Ratio) (Paquot, 2019). Phraseological sophistication was operationalized as the proportion of each type of unit on the Academic Collocation List (Ackermann & Chen, 2013) as well as the mean PMI of the units based on the L2 Research Corpus, a corpus of published research papers in Applied Linguistics (Paquot, 2018, 2019) as well as the proportion of units in six bands of varying PMI strength (Paquot, 2018). As a means of comparison, Paquot also calculated several commonly-used measures of lexical and syntactic complexity.

The results showed that phraseological sophistication increased significantly with proficiency and that two measures (mean PMI of adjectival modifiers and direct objects)[6] were significant predictors of the holistic ratings in a mixed effects regression model, explaining 25% of the variance in CEFR scores. In the first study, none of the lexical or syntactic measures were found to differ significantly across proficiency levels (Paquot, 2019) and in the follow-up study (Paquot, 2018) only one measure of lexical complexity (lexical density) had a significant effect of proficiency but this measure was not found to significantly improve model fit during the stepwise model selection process. Taken together, this seemed to suggest a relationship between the construct of phraseological complexity and L2 proficiency (at least for the sophistication of adjectival modifier, adverbial modifier and direct object dependency relations). Moreover, it showed that for these highly advanced learners, measures of phraseological complexity seemed to be more informative about their proficiency than the lexical or syntactic complexity measures that Paquot employed. The following section reviews other studies that have investigated the relationship of phraseological complexity with L2 proficiency and development. It is important to note that although these studies often did not use the term *phraseological complexity* themselves, they have used measures that fit Paquot's (2019) definition of the construct (i.e., the diversity and sophistication of phraseological units) and therefore contribute to investigating the usefulness of such measures as indices of proficiency and development.

2.2.2 Phraseological complexity in L2 writing

Diversity. Although Paquot (2019) did not find any significant association between phraseological diversity and proficiency, other studies have found an effect for diversity measures. Using a similar methodology to that of Paquot (2019), Jiang et al. (2021) found that the diversity (RTTR) of both adjectival modifier and direct object dependency relations were significant predictors of the scores attributed to low-

[6] Measures based on the proportion of units from the Academic Collocation List did increase somewhat systematically with proficiency but the differences were not found to be significant.

intermediate (CEFR B1) level written narratives produced by Chinese learners of English. Another study that investigated phraseological diversity was that of Garner (2020), who extracted verb + noun collocations from TOEFL argumentative essays rated as low, intermediate and advanced. In this study, diversity was operationalized as the normalized entropy between collocation tokens and collocation types in each text. Used in this way, normalized entropy represents the "distribution of collocation types among collocation tokens" (Garner, 2020: 28) and thus gives an indication of whether a text is composed of a few unique types or many different repeated types, which constitutes diversity. This measure is also less sensitive to the Zipfian distributions that are typical of TTR or RTTR. The results showed that a one standard deviation increase in entropy was associated with a 19.90% increase in the odds that a given text would be rated as more proficient. Rubin, Housen and Paquot (2021) also used a corpus of proficiency exams, but for L2 learners of Dutch. They found that a one unit increase in the diversity (RTTR) of adverbial modifiers (adverb + verb/adjective/adverb) was associated with a 245% increase in the odds of passing a CEFR B1 level proficiency exam and a 54% increase in the odds of passing a B2 level proficiency exam. The fact that these studies used more beginner-level learners compared to the participants in Paquot (2019) may indicate that phraseological diversity measures are more predictive of lower levels of proficiency. Unfortunately, these are the only studies that have measured phraseological diversity in L2 writing so these conclusions must remain tentative.

Sophistication. Phraseological sophistication on the other hand, has received much more attention. For example, the studies mentioned above found that the PMI of phraseological units was a significant predictor of proficiency (in the case of Jiang et al., 2021) and the odds of passing a proficiency exam (Rubin et al., 2021). One reason that sophistication has received more attention than diversity is likely due to the availability and widespread use of the Tool for the Automatic Analysis of Lexical Sophistication (TAALES, Kyle & Crossley, 2015). Despite the name, the tool actually calculates several measures which would probably more appropriately be called measures of phraseological sophistication. These include measures of the PMI of bigrams and tri-

grams (continuous sequences of two/three words) in large L1 English reference corpora. Various studies using this tool have reported significant effects of PMI on predictions of L2 proficiency (Bestgen, 2017; Garner et al., 2018, 2019; Kim et al., 2018; Zhang & Li, 2021).

The strong relationship between phraseological sophistication (PMI) and proficiency reported in cross-sectional studies has also been born out to some extent in longitudinal studies. For example, Siyanova-Chanturia (2015) found that beginner-level learners of English significantly increased the proportion of adjective + noun collocations with a PMI greater than 3 used in narrative tasks collected over a period of five months.[7] Bestgen and Granger (2018) likewise showed that A2-C2 level learners increased the proportion of high PMI bigrams used in argumentative essays over a period of three years. In a study using the LANGSNAP data, Edmonds and Gudmestad (2021) found that the mean PMI of adjective + noun collocations that the learners of French used in argumentative essays increased significantly over the course of 21 months. However, other studies report no significant change in PMI over a shorter time period (i.e., one semester in Bestgen & Granger, 2014; Yoon, 2016). In at least one study, PMI was actually found to decrease over time. Siyanova-Chanturia ad Spina (2020) measured the PMI of noun + adjective collocations used in narrative tasks by A1-B2 level Chinese learners of Italian. Somewhat counter to expectations, the PMI of the units significantly decreased over the course of six months. This seems to support the claim by N. C. Ellis et al. (2008) that knowledge of association strength develops very slowly, only after learners are exposed to a vast amount of target-language input. It also suggests that the development of phraseological sophistication may not be linear. Learners may initially increase their use of highly frequent (but low PMI) phraseological units and then as they receive more target-language input, begin to rely more on highly exclusive (high PMI) phraseological units. Such long term patterns may only be visible when comparing across wide ranges of proficiency and not when tracing a more homogeneous group of learners over a short time period. This interpretation is supported by the results of Paquot,

[7] A PMI of 3 is commonly considered a threshold for collocability (Siyanova & Schmitt, 2008).

Naets and Gries (2021) who compared the effects of proficiency scores and development over time, and found that proficiency of the learner (the results of their Oxford Quick Placement Tests) explained more variance in the PMI of direct object dependencies than time in years. To sum up, these studies suggest that, in general, the diversity and sophistication of phraseological units learner productions can be useful indicators of proficiency and development in L2 writing but that the strength of this relationship may not be the same for beginner and advanced proficiency levels.

2.2.3 Phraseological complexity in L2 speech

Diversity. There are only two studies that have looked at phraseological diversity in L2 speech. One such study was conducted by Paquot et al. (in press) who compared the phraseological diversity of B1- to C2-level oral exams from the Trinity Lancaster Corpus (TLC, Gablasova et al., 2019). The results showed that there was a significant increase in the diversity (RTTR) of words used to fill verb + direct object structures in conversational tasks between the B2 and C1 levels. The only other study that has looked the diversity of phraseological units in L2 speech is that of Qi and Ding (2011) who found a significant increase in the diversity (TTR) of manually-identified formulaic sequences used in the oral monologues produced by Chinese learners of English over the course of a four-year bachelors program. While these results are promising, in that they seem to be in line with the findings for phraseological diversity in writing, there is clearly a need for more research investigating phraseological diversity in L2 speech, especially in light of the differences between written and oral production (see Section 2.3).

Sophistication. In terms of sophistication, while some cross-sectional studies have reported a significant positive relationship between PMI and oral proficiency (Eguchi & Kyle, 2020; Uchihara et al., 2021; Zhang et al., 2021), the results are less consistent than for writing. Paquot et al. (in press), for example, found that the PMI of direct objects in TLC oral exam productions actually decreased from B1 to B2. Likewise, Tavakoli and Uchihara (2020) found a decrease in the PMI of bigrams used in

argumentative monologues from B2 to C1. Just as for writing, the longitudinal development of phraseological sophistication in speech has been found to be non-significant over time: four months in the case of Garner and Crossley (2018) and one year in the case of Kim et al. (2018) (both studies included a broad range of proficiency levels).

Taken together, these results suggest that the relationship of phraseological complexity to proficiency may not be the same across modes. This may be because learners are more likely to take risks in speech as compared to writing, trying out more sophisticated vocabulary in different (but non-conventional) word combinations. It could also be that in the pressured condition of online production, beginner learners rely on a smaller set of phraseological teddy bears (as in Nesselhauf, 2005), which raises the average PMI of the phraseological units relative to the more advanced learners who are more open to experimentation. To date, only one study has directly compared the phraseology of oral and written L2 production however. Using a corpus of TOEFL proficiency exams, Biber and Gray (2013) compared oral and written responses to integrated and independent TOEFL exam tasks. The authors extracted frequent collocates[8] for five highly-frequent verbs (*get, give, have, make* and *take*). The results showed that the use of these frequent collocations was highest in oral, integrated tasks at the intermediate level. The authors interpreted this finding both in terms of modality differences and proficiency:

> "...these patterns suggest that collocation sequences are acquired at intermediate levels and that their use requires some planning and processing time but, at the highest levels and in the tasks with the most opportunity for planning and production, they are less commonly used, possibly because they are stigmatized as being clichés and less creative." (Biber & Gray, 2013: 31).

In other words, decreases in PMI at intermediate levels of oral-

[8]The authors operationalized collocation in a somewhat different way than most of the studies cited in this section. In this study, collocation was defined as follows: "any content word that co-occurred with the target verb in more than ten texts at a rate of more than five times per 100,000 words." (Biber & Gray, 2013: 19). These collocates were therefore based on frequency within the learner texts themselves rather than a separate L1 reference corpus.

proficiency (as in Paquot et al., in press; Tavakoli & Uchihara, 2020) may result from the production pressures of speech. This would be in line with the psycholinguistic research showing that frequent phraseological units are the ones that are most easily accessed by learners (N. C. Ellis et al., 2008), an effect which may be even more evident in the case of pressured online production of speech. Only as learners are exposed to more target-language input, do they begin to produce more strongly associated phraseological units but the results of Biber et al. (2013) suggest that modality may continue to play a moderating role in the ability of learners to access and produce such units. This will be revisited in Section 2.3.

As shown by the studies discussed in this section, although there is some inconsistency between studies, there is good reason to think that, in general, phraseological complexity increases with proficiency and development in L2 English. It is striking however, that in the only study looking at a non-Germanic language (L2 Italian), Siyanova-Chanturia and Spina (2020) found a decrease of phraseological complexity over time. With this in mind, the following section turns to studies of phraseology in L2 French. As will be evident, research in L2 French has approached phraseology from another perspective and has used different methodologies to the studies discussed in the previous section. Nevertheless, there are important parallels that can be drawn with the *phraseological complexity* literature discussed above.

2.2.4 Phraseology in L2 French

Early research into the phraseology of L2 French focused on the contribution of continuous and discontinuous word combinations (e.g., *il y a*, 'there is'; *c'est...que*; 'it's ... that') to the development of fluency (Raupach, 1984; Towell et al., 1996), finding that learners increased their use of such expressions following time spent abroad but that compared to native speakers, they tended to overuse a restricted set of them, resulting in "non-idiomatic performance." More recently, Bolly (2008, 2009), compared verb + noun collocations with the verbs *prendre* ('to take') and *donner* ('to give') between L1 and L2 argumentative essays.

Compared to the L1 group, Bolly found that the learners under-used collocations with *prendre* and over-used collocations with *donner*. For example, in the L1 data, there were 34 instances of *prendre + exemple* ('take + example') compared to only one such instance in the learner data. This suggested that learners had not yet acquired native-like knowledge of the distributional patterns governing the use of these verb-noun collocations.

Although there are some exceptions (e.g., the study of Edmonds & Gudmestad, 2021 discussed in Section 2.2.2), the majority of phraseology research in L2 French has focused on the use of *formulaic sequences* (FS)[9], defined as "the conventionalized combination of at least two graphic words favoured by native speakers in preference to (an) alternative combination(s) which could have been equivalent had there been no conventionalization." (Erman & Warren, 2000: 31). On a practical level, these units are not identified using statistical criteria (i.e., frequency of occurrence greater than expected on the basis of chance) but rather are identified manually using the criteria of *restricted exchangeability*: "at least one member … cannot be replaced by a synonymous item without causing a change of meaning or function and/or idiomaticity" (Erman & Warren, 2000: 32). This is typically verified by comparing web-search results for the sequence in question and a synonymous sequence (Forsberg, 2008, 2010). Any sequences that are at least twice as frequent compared to the alternative are considered conventional. As an illustration, Forsberg (2010: 37) describes how *gestion de portefeuilles* ('portfolio management') returns 184 000 hits on Google, compared to *administration de portefeuilles* ('portfolio administration') which only yields 119. This method can be considered a more top-down approach to phraseology (see Granger & Paquot, 2008), in comparison to the more bottom-up statistical approach used by Paquot (2019). Nonetheless, there is some overlap between the formulaic sequences identified by this method and the *phraseological units* extracted through more automated

[9]There is some inconsistency in the literature regarding the use of the this term. Erman and Warren (2000) actually use the term 'multi-word unit' whereas Forsberg (2008) uses the term 'conventionalized sequence.' However, more recent publications (e.g., Forsberg Lundell et al., 2014; Forsberg Lundell & Lindqvist, 2014) use the term 'formulaic sequence' so I have followed that usage here.

methods. For instance, Forsberg (2010) gives the example of *avoir envie* ('to want'), which would also be identified as a direct object dependency relation (as in Paquot, 2018, 2019), a contiguous bigram (as in Granger & Bestgen, 2014; Kyle & Crossley, 2015) or a verb + noun collocation found through a regular expression search (as in Garner, 2020). This means that despite the differences in methodology, it is nonetheless possible to draw comparisons between results of studies using this method and the results of the more statistically-based studies reviewed in Section 2.2.2 and 2.2.3.

There are, for example, clear links between the construct of phraseological complexity and studies conducted by Forsberg (2008, 2010), who compared the quantity and diversity of lexical FS (those sequences with a "denotative or referential meaning") in interviews conducted with Swedish learners of French with varying amounts of target-language exposure: beginners with only one month of classes, university-level learners, high school students and a group of very advanced learners who were long term residents in France as well as two native speaker control groups (Erasmus students at the Swedish university and residents of Paris). The results showed that the number of lexical FS and the diversity of such sequences (TTR) increased according to the amount of exposure each group had to French but that the very advanced group was not significantly different from the native speakers. Similar results were also reported by Forsberg Lundell and Lindqvist (2014), who found a difference in the number of lexical FS used by university students compared to long term residents in France but no difference between the long term residents and native speakers. It is worth noting however that studies using tasks other than interviews *have* found differences between the advanced learner group and native speakers in terms of lexical FS use. For example, Forsberg and Fant (2010) used a corpus of role plays (telephone calls with a superior) and narratives (retelling of a video). The advanced learners used significantly fewer phrase-based lexical FS[10] overall compared to native speakers and used

[10] Phrase-based lexical FS are defined as those that represent complete phrasal structures (i.e., NPs, VPs, AdjPs, AdvPs, PrepPs) such as *faire du sport* ('practice a sport') as opposed to those that comprise a full clause (e.g., *je vous remercie de votre compréhension*, 'thanks for being so understanding') (Forsberg & Fant, 2010: 59).

fewer phrase-based lexical FS in the role play than in the narrative task. The authors attributed these differences both to the need for specific "denotative-referential lexis" in the role play task as well as to the "degree of difficulty of the tasks involved" (Forsberg & Fant, 2010: 63). In another study using the same corpus, Erman et al. (2015) additionally found that learners had significantly less diversity of lexical FS than native speakers in the narrative task but not in the role play task, again attesting to the role that task-based characteristics may play regarding L2 phraseology (see further discussion of this issue in Section 2.3.4).

In terms of sophistication, while Forsberg (2010) did not directly compare the association strength of sequences used by L1 and L2 speakers, she did calculate the PMI of the ten most frequently used sequences by the advanced learner and L1 groups using a small L1 reference corpus. The average PMI of these units was found to be 10.3, indicating that these groups used relatively sophisticated lexical FS. In addition to the findings for spoken production, Forsberg and Bartning (2010) also found that the number of lexical FS significantly increased with proficiency (CEFR) in written productions. In light of these findings, Forsberg Lundell et al. (2014) argue that phraseology may be more indicative of advanced proficiency than syntactic and morphological features (cf. Bartning & Schlyter, 2004), which would be in line with the findings of Paquot (2018, 2019) for L2 English. Taken together, these studies suggest that, as in English, phraseological complexity may be informative about proficiency and development in L2 French. However, no study has yet directly investigated phraseological complexity (in terms of diversity and sophistication) in L2 French. This constitutes an important gap that this thesis aims to fill.

2.3 Complexity and modality

Writing and speaking are typically claimed to differ in three main ways: the presence/absence of an audience, the stability of the language signal and the degree of control that a user has over linguistic output (Ravid & Tolchinsky, 2002). The confluence of these factors means that written and spoken output are often very different. In the case of

French more specifically, Duneton (1973: 179) notes:

> "...[l]e Français moyen d'aujourd'hui se trouve dans une situation hautement bizarre : il n'écrit pas la langue qu'il parle et il ne parle pas la langue qu'il écrit."
>
> *...[t]he average Frenchman today finds himself in a very bizarre situation: he doesn't write the language that he speaks and he doesn't speak the language that he writes*.

Putting aside this bit of hyperbole, Duneton does touch on something quite important. Because of the differences in production between the two modes, writing and speech also tend to be functionally quite different in terms of the kind of information they express and the types of linguistic features used to express that information. The following section explores the issue of modality in more detail to show how L2 complexity has typically been thought to differ between writing and speaking.

First, in order to understand the basis for such differences, Section 2.3.1 provides a brief overview of the most commonly accepted cognitive models of oral and written production. The purpose of this section is not to offer an in-depth critique of these models (which would be beyond the scope of this thesis) but rather to provide the background that is relevant for understanding the theoretical underpinning of cross-modal L2 complexity studies. Section 2.3.2, then shows how modality has typically been conceptualized within the tradition of Task-based Language Teaching (TBLT) and reviews the handful of studies that have directly compared complexity between written and oral L2 production. Finally, Section 2.3.3, explores how looking at modality as a type of register variation (Biber, 1988; Biber et al., 1999) may help to explain some of the inconsistencies reported in studies that have tried to compare L2 complexity across modes in the TBLT-tradition.

2.3.1 Production models

Spoken production. One of the most influential models of speech production is that of Levelt (1989, 1999). A major tenet of this model is

that speech production is modular and involves three separate components, namely the conceptualizer, the formulator and the articulator. As shown in Figure 2.4, the conceptualizer is responsible for generating the pre-verbal message, drawing on social competence and knowledge of the world to generate a conceptual structure composed of lexical concepts. In the formulator, the pre-verbal message is translated to linguistic form by linking the conceptual structure to the corresponding lemmas in the mental lexicon. The mental lexicon contains rich information about the semantic properties for each lemma, corresponding syntactic frames, collocational relationships between lexical items in the network as well as the corresponding phonological and articulatory characteristics. In this model, speech production is claimed to be lexically-driven, that is, the selection of a lemma leads to the activation of corresponding syntactic structures through a mechanism known as *activation spreading*. Signals are sent between neurons, which causes activation to spread from one item to another. Decisions are made once a given node (representing, for example, a concept or a word form) reaches a threshold level of activation (Hebb, 1949). The result of this process is a surface structure that undergoes morpho-phonological encoding, producing a phonological score. Each syllable in the phonological score then undergoes phonetic encoding in order to link it to a corresponding articulatory gesture. This articulatory score is then passed to the articulator, which activates the appropriate laryngeal and supra-laryngeal movements to produce speech. The output of both internal speech (the phonological score) and external speech are monitored and corrected if necessary in a feedback loop.

An important aspect of this model is that the three components are believed to operate incrementally (Fry, 1969; Garrett, 1976; Kempen & Hoenkamp, 1987) that is, information is passed from one component to the next. Processing is triggered in subsequent components as soon as they receive a fragment of input from the previous one. This means that the three components can operate largely in parallel (beginning on a new fragment as soon as the previous one has been transferred on). In addition, whereas conceptualization requires conscious attention, formulation and articulation proceed largely au-

tomatically, which accounts for why speech is able to be produced relatively quickly and smoothly, at least in the case of L1 speech production.

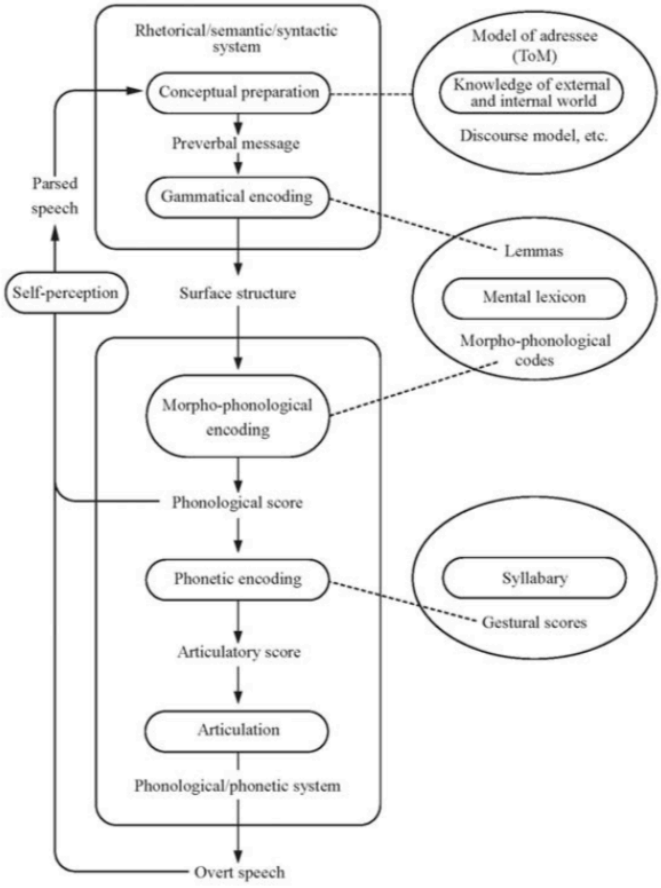


Figure 2.4: A blueprint of the speaker (Levelt, 1999: 6)

Many researchers agree that the general psycholinguistic mechanisms in L2 production are largely the same as those proposed by Levelt (1989, 1999) for L1 production (De Bot, 1992; Kormos, 2006; Poulisse & Bongaerts, 1994; Towell et al., 1996). Just as in L1 production, L2 production is theorized to contain separate modules for conceptualizing, formulating and articulating linguistic units (Figure 2.5).

The three modules are also believed to work incrementally, allowing for separate fragments to be processed in subsequent modules as soon as they are encoded by the previous module. However, in L2 production, the lexicon may not contain all entries required for the pre-verbal message or the entries themselves may not contain enough information (e.g., about semantic properties) to meet the demands of the conceptualizer. As argued by Towell (2012: 55), this has several implications for the complexity of the interlanguage system:

> "The degree of elaboration associated with an aspect of syntax is likely to be associated with the extent to which the learner has been able to create a full syntactic tree in his or her interlanguage. Lexical elaboration will relate to the extent to which lexical items have been practised successfully in a variety of contexts. For both syntax and lexis, the degree of depth of knowledge and the stability of that knowledge will relate to the extent to which procedures have been laid down in memory for the processing of the syntax and lexis, allowing always for the possibility that non-native like knowledge (intermediate interlanguage knowledge) may also acquire depth and stability and lead to fossilization."

In addition, for the L2 learner, the network of connections between entries may be less organized, which means that collocational networks are less elaborate and that the syntactic frames corresponding to lexical entries are not easily accessed, leading speakers to draw on declarative memory for certain syntactic rules. All of this can cause difficulty at the formulator stage, requiring more conscious attention on behalf of the learners (Skehan, 2009b) and impairing the ability for parallel processing (Kormos, 2006).

Written Production. While there are several proposed models of written production (e.g., Börner, 1987; Flower & Hayes, 1980; Galbraith, 2009; Hayes, 2012), the discussion here focuses on Kellogg's (1996) model. As argued by Michel et al. (2020), this model emphasizes the linguistic encoding process (e.g., as opposed to socio-cultural factors), which makes it particularly useful in explaining L2 writing. In Kellogg's

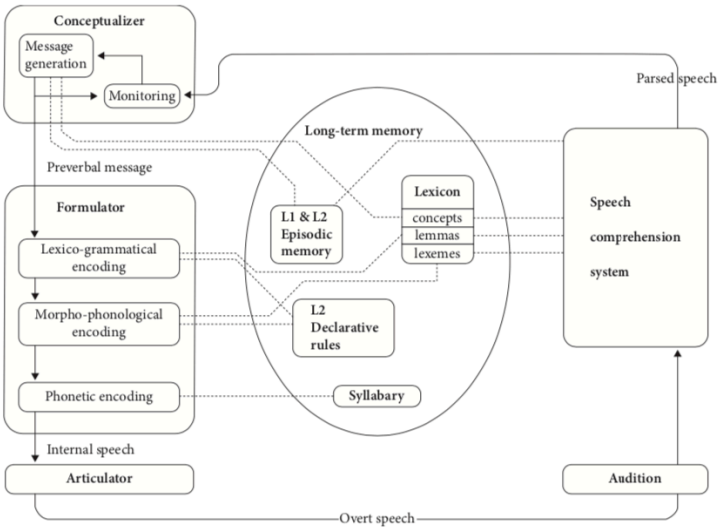


Figure 2.5: A model of L2 speech (Kormos, 2011: 43)

model, writing is also considered to be a modular process, containing three main components: formulation, execution and monitoring (Kellogg, 1996). The formulation stage involves processes related to Levelt's conceptualizer and formulator. At this stage, a writer creates a plan, the result of which is a propositional representation which is then translated into linguistic units.[11] Often this is done via an intermediate *partial translation* which allows a writer to "plan and tentatively translate ideas before beginning the first draft" (Kellogg, 1996: 9). Ostensibly, this uses the same mechanism as in speech (i.e., lexically-driven activation spreading). At the execution stage, this translation is mapped onto the motor functions required for handwriting or typing.

Throughout the process, the writer is able to monitor output and revise when necessary. Editing can happen both before the text is executed (at the 'partial translation' stage) as well as after the text has

[11] It is important to note that Kellogg's use of the word 'plan' is not meant to imply that writers literally draw up a point-by-point physical outline on paper every time they write but rather, this should be seen more akin to the pre-verbal output of the conceptualizer in Levelt's (1989, 1999) model.

been produced. This means that in comparison to speech, writing is prototypically a cyclical, recursive process. In addition, while these three components may occur simultaneously (provided the execution process is somewhat automatized), in general, writers are able to spend more time planning prior to the execution stage, freeing up attentional resources needed for the cognitively demanding process of translating ideas into linguistic units. This is particularly important to keep in mind when considering the differences in complexity between oral and written L2 production, which is the topic of the following section.

2.3.2 Complexity and Modality in Task-based Language Teaching

In Task-based Language Teaching (TBLT), modality is seen as one of many task related factors influencing the attention that a learner is able to pay to linguistic output (R. Ellis, 2003; Robinson, 2015; Skehan, 1996, 1998). Within this tradition, there are two models to explain the influence of such factors on L2 performance. The Limited Attentional Capacity (LAC) model (Skehan, 2014) claims that learners have limited attentional resources and as such, there is often a "trade-off" between complexity, accuracy and fluency such that higher performance along one dimension may be associated with lower performance in another dimension. This stands in contrast to the Cognition Hypothesis (Robinson, 1995, 2001, 2011, 2015), which claims that attentional resources can expand to meet functional demands provided that tasks are appropriately sequenced and designed to avoid having to draw on the same "resource pools." As N. C. Ellis & Wulff (2019) point out, these two models make similar predictions about the way that modality should affect complexity. For a beginner L2 learner, the limited state of the L2 mental lexicon means that the formulator may not be able to completely meet the demands of the conceptualizer. As a result, the learner may need to find an alternative way to express what they want to say, derailing the normally smooth process of language production. According to the LAC model, when a task is completed in the written mode, it allows for more online planning (Manchón, 2014), thereby easing the burden on the formulator and leading to higher levels of complexity

Table 2.1: Previous cross-modal studies of L2 complexity

Study	Task Type	Lexical Diversity	Lexical Sophistication	Syntactic Complexity (Clause)	Syntactic Complexity (Noun Phrase)
R. Ellis and Yuan (2005)	Narrative	+		+/=†	
Granfeldt (2007)	Narrative/Expository	+		=	
Kormos and Trebits (2012)*	Narrative	+		=	
Kormos (2014)*	Narrative	+	+	=	+
Kuiken & Vedder (2011)	Argumentative	+		+	
Zalbidea (2017)	Argumentative	+		-	
Vasylets, Gilabert & Manchón (2017)	Argumentative	+	=	+	=
Vasylets, Gilabert & Manchón (2019)	Argumentative	+	+	+	+

Note:
+ Advantage for writing, - Advantage for speaking, = No significant difference between writing and speaking
* Based on the same data set.
† Significantly higher in the 'pressured' condition but not in the 'careful' condition.

in the output. Similarly, under the CH, modality can be seen as an element that can either disperse attentional resources, for example, by forcing speakers to attend to conceptualization and formulation simultaneously; or as an element that can direct attentional resources, for example, by allowing writers to pay more attention to linguistic form during the editing and revision process (Kormos, 2014). In this way, both the LAC and CH models claim that L2 writing should exhibit higher levels of complexity as compared to speech. As it turns out however, this general prediction has not been clearly born out in empirical research.

In order to investigate the effects of modality, most TLBT studies attempt to control for task differences by having learners complete the same or very similar tasks in both modes. As shown in Table 2.1, this is usually either a narrative-type task or an argumentative-type task. These are each discussed briefly below.

Studies using narrative-type tasks. One of the first studies to directly compare complexity across modes using a narrative-type task was that of R. Ellis and Yuan (2005), who used a picture-based task, to compare the complexity in oral and written productions elicited from university-level Chinese learners of English. In addition to completing the task in both modalities, the learners were randomly assigned to either a *pressured condition*, where a time limit was imposed on

the task (5 minutes for the written task and 17 minutes for the oral task) or a *careful condition*, where they had unlimited time to complete the task.[12] The results of the study revealed that when time to complete the task was constrained (i.e., the pressured condition), written productions had higher levels of syntactic complexity (clauses per T-unit) and were more lexically diverse (MSTTR) as compared to the oral productions. When participants had unlimited time to complete the task however (i.e., the careful condition), written productions were more lexically diverse than the oral productions but not significantly different in terms of syntactic complexity. According to the authors, this suggests that it is primarily the opportunity for online planning that leads to higher complexity in writing as compared to speaking: "writing, even when pressured, allows more time for planning, formulating, executing and monitoring messages than speaking" (R. Ellis & Yuan, 2005: 189). Another study that directly compared complexity in oral and written narrative tasks was Granfeldt (2007). In this study, Swedish learners of French watched a short video and completed a narrative and expository task in both modes, this time without time constraints (as in the 'careful' condition in the R. Ellis and Yuan study). In line with the results of R. Ellis and Yuan (2005), the written productions[13] were found to be more lexically diverse (D) but there was no significant difference in syntactic complexity (clauses per T-unit) across modes. Similar results are also reported by Kormos (2014) and Kormos and Trebits (2012) who looked at Hungarian learners of English completing cartoon- and picture-based narrative tasks. Written productions had higher lexical diversity (D) and sophistication (proportion of low-frequency words) but there were no significant differences across modes for syntactic complexity (operationalized as the subordinate clause ratio). In contrast to the previously discussed studies however, this study also measured the number of modifiers per noun phrase, and this measure *was* found to be significantly higher in writing as compared to speaking.

[12] Following standard practice in TBLT studies of planning, the time limit for the written pressured condition was set on the basis of a pilot study and reflected the time that the fastest learner took to complete the task. The authors do not provide a justification for the choice of the time limit in the oral pressured condition.

[13] In this analysis, Granfeldt (2007) collapsed both task types together and the differences (if any) between the two task types is not reported.

Taken together, these studies suggest that, in line with the predictions of the LAC and CH models, the written mode does seem to promote higher levels of lexical complexity. The picture for syntactic complexity is less clear cut with most studies finding no significant differences across modes at least for clause-level complexity. Syntactic complexity at the phrase level however may be more informative for cross-modal differences than complexity at the clause level (Kormos, 2014; Kormos & Trebits, 2012), which is in line with the literature on register variation as discussed in Section 2.3.3.

Studies using argumentative-type tasks. A second set of studies have looked at cross-modal complexity differences using argumentative-type tasks. Kuiken and Vedder (2011), for example, asked Dutch learners of Italian to offer advice (either orally or in written form) to a friend about the choice of a holiday destination. Written productions were found, on average, to have slightly higher levels of lexical diversity (CTTR) and to be more syntactically complex (clauses per T/AS-Unit, dependent clauses per clause). Unfortunately, the authors did not directly test the differences across modes, so it is not clear whether these differences would have been statistically significant. Using a similar task (arguing for the choice of a bed and breakfast), Zalbidea (2017) compared oral and written productions by English-speaking learners of Spanish. In line with the previously discussed studies, the written productions were found to exhibit more lexical diversity (RTTR) than oral productions. However, this study is somewhat unique in that it is one of the only studies to report higher levels of syntactic complexity (subordination) in speech as compared to writing (but see also Ferrari & Nuzzo, 2009), and shows that cross-modal differences in syntactic complexity are somewhat less clear-cut than for lexical complexity, where there is a clear tendency for higher complexity in the written mode. Another study to use more of an argumentative-style task was that of Vasylets, Gilbert and Manchón (2017) who compared oral and written productions by Spanish/Catalan-speaking learners of English completing the Fire Chief Task (Gilabert, 2007). In this task, learners need to explain how to rescue people from a burning building while considering factors such as the number of people, the danger faced by the different 'victims'

etc. In terms of syntactic complexity, written productions exhibited higher levels of complexity at the clause-level (longer AS-units and AS-units with a higher proportion of clauses) compared to spoken productions. Complexity at the phrase level (number of modifiers per noun phrase) was not significantly different across modes however. Lexical diversity (D) was also higher in writing as compared to speaking but lexical sophistication[14] was not found to differ significantly across modes. A more recent study by the same authors, using a larger sample of learners, found that written productions exhibited higher levels of complexity across all lexical (density, variety, richness and sophistication) and syntactic (coordination, subordination and nominal) complexity measures (Vasylets et al., 2019).

As shown in Table 2.1, the influence of modality seems much stronger for lexical complexity than for syntactic complexity. In almost all studies comparing complexity across modes, lexical complexity was found to be higher in written as compared to spoken productions.[15] In the case of syntactic complexity, studies looking at narrative-type tasks have found little to no difference across modes for clausal complexity. Studies looking at argumentative tasks however, have found significant differences across modes but the direction of the effect is not consistent with Zalbidea (2017) finding higher levels of subordination in the speech produced by her English learners of Spanish but Vasylets et al. (2017, 2019) finding higher levels of subordination in the written productions of their Spanish/Catalan learners of English. Of course, these studies focused on different target languages and the tasks completed by the learners were not exactly the same, so it is difficult to directly compare the results. Perhaps more problematically, these studies, as well as all of the studies mentioned above operationalized syntactic complexity using omnibus measures such as the length of syntactic constituents or subordination ratios, which means

[14]The lexical sophistication measure used in this study was based on a formula which calculated the proportion of tokens in 1000- and 2000-word frequency bands as well as the proportion of tokens from the Academic Word List (Coxhead, 2000) and those off-list and divided that by square root of the number of tokens. The authors do not mention which reference corpus was used for the frequency bands.

[15]One exception to this is Yu (2010), who found no significant differences in the lexical diversity between written compositions and oral interviews. However, in this study L1 and L2 data were collapsed together which makes it difficult to compare the results with the other studies outlined above.

that meaningful differences in the syntactic structures used in speech versus writing may have been collapsed (Biber et al., 2011, 2020) as will be discussed in the following section.

2.3.3 Register-based approaches to modality

As argued by Halliday (1989, 2004), speech and writing constitute fundamentally different ways of constructing discourse. Speech tends to present a dynamic view of the world, focusing on actions and processes whereas writing tends to present a synoptic view, focusing on things that exist in the world. As a result, "spoken language favours the clause, where processes take place, whereas written language favours the nominal group, the locus of the constitution of things" (Halliday, 1989: 99). This difference is illustrated in Examples 2.2 and 2.3. Although they both provide the same information, the written example is much more informationally dense, compacting the equivalent of five clauses into just one.

Example 2.2. Writing (Halliday, 1989: 79):

> "The use of this method of control unquestionably leads to safer and faster train running in the most adverse weather conditions."

Example 2.3. Speech (Halliday, 1989: 79):

> "You can control the trains this way and if you do that you can be quite sure that they'll be able to run more safely and more quickly than they would otherwise no matter how bad the weather gets."

Starting in the 1980s, Biber (1985, 1986) began to empirically investigate these cross-modal differences by comparing the linguistic features found in large oral and written corpora using a methodology called Multidimensional Analysis (MD). Broadly speaking, this technique uses statistical factor analysis to discover linguistic features that co-occur in different oral and written registers. MD studies have consistently shown that a characteristic of oral texts in English is that they tend to be associated with a greater use of pronouns, modal verbs, lexical verbs and finite clauses (Biber, 1988; Biber et al., 1999).

Functionally, this is associated with the more personal or involved nature of oral communication and in particular with the expression of stance which requires personal pronouns and dependent clauses (e.g., "I think that...," "I guess that..."). Written texts on the other hand, tend to be marked by the use of nouns, nominalizations, attributive adjectives and prepositional phrases because these linguistic features allow writers to convey more information in a limited space. Of course, there is a large variation in the extent to which different registers reflect these linguistic characteristics, but this tends to go in hand with the extent to which different registers have a personal or involved versus informational purpose. The more personal the communicative function of a text (e.g., personal letters vs. research articles), the more a text tends to have *oral*-like linguistic features. Another key insight from the MD-based research is that there is much more variation within written registers than within oral registers (Biber, 1992). In other words, most oral texts are characterized by a clausal-focus but the degree to which a written text has a clausal rather than a phrasal focus depends largely on its communicative function. These patterns have been attested in various registers as well as in many different languages (Biber, 2014).

Although there have not been MD studies of French specifically, there is evidence that similar patterns exist. Blanche-Benveniste (1995), for example, compared coverage of events in newspapers and oral-retellings. She found that oral retellings used more finite verbs and subordination whereas newspaper accounts tended to use one main verb and made use of participles and infinitives. This can be seen in Example 2.4, which tells the story of a car accident. The goal of the speaker is to describe the sequence of actions surrounding the event and to provide explanation. As such, the text is organized around the finite verbs describing the process (e.g., *se sont mis aux fenêtres*, 'went to the windows'; *a fait un grand bruit*, 'made a large noise'; *les pompiers sont arrivés*; 'the firefighters arrived' etc.) as well as subordinate clauses providing the interpretation of the speaker (e.g., *parce que ça m'avait choquée*, 'because it shocked me'; *puisque c'était normal*; 'because it was normal').

Example 2.4. Oral retelling (Blanche-Benveniste, 1995: 25):

> "les gens se se sont mis aux fenêtres puisque ça a fait un grand bruit la voiture était en accordéon surtout à l'avant et euh ils ont appelé /les, des/ pompiers - les pompiers sont arrives - moi j'ai fait une crise de nerfs parce que ça m'avait choquée - ma soeur était dans un sale état aussi et euh les pompiers sont arrivés donc et ils nous ont menés à l'hôpital alors euh en plus que nous avions eu un accident de voiture les pompiers n'ont rien trouvé de mieux que de rire - mon père s'est énervé puisque c'était normal et euh il y a eu une petite dispute en plus de ça."
>
> *the people all gathered around the windows because there was a loud noise the car was [squished like] an accordion especially at the front et uh they called the some firefighters - the firefighters arrived - me I had a little breakdown because it shocked me - my sister was in a bad state also and the firefighters arrived and so and they took us to a hospital so uh also not only did we have a car accident but the firefighters also couldn't do anything but laugh - my father was annoyed because it was normal and there was a little argument as well.*

Example 2.5, which comes from a newspaper story, is much more to-the-point, with the entire sequence of events centered around one finite verb (*ont été pendus*, 'were hanged'). Additional information is communicated not through subordinate clauses but rather through the use of past participles (e.g., *condamnés à mort*, 'condemned to death') and nominalizations (e.g., *détention d'armes*, 'arms possession').

Example 2.5. Newspaper account (Blanche-Benveniste, 1995: 27) :

> "Condamnés à mort par la justice militaire pour avoir adheré à des organisations clandestines, ainsi que pour détention d'armes et d'explosifs dans le but de renverser le régime, trois islamistes ont été pendus jeudi 16 décembre, au Caire."
>
> *Condemned to death by the military tribunal for having*

> *belonged to clandestine organizations, as well as for the possession of weapons and explosives with the intention of overthrowing the regime, three islamists were hanged Thursday, December 16th in Cairo.*

Along the same lines, Labbé (2007) compared a corpus of oral interviews with a corpus of literary French. The oral interviews were characterized as having a higher proportion of verbs, especially past participle and finite verbs and exhibited more subordination than the written corpus. The literary corpus on the other hand, had a higher proportion of common nouns, adjectives, present participles and infinitives. This seems to suggest that, in line with the findings for English and other languages (Biber, 2014), French also seems to prefer a more clausal-focused discourse style in speech and a more phrasal/nominal-focused discourse style in writing.

As argued by Biber et al. (2011, 2020, 2021), a major implication of these corpus-based findings is that the omnibus measures of syntactic complexity (e.g., subordination ratios) that are frequently used in CAF/TBLT-based SLA research are too coarse-grained (an argument also made by Norris & Ortega, 2009). Among other problems, omnibus measures do not distinguish between different functions of subordinated clauses (e.g., adverbial clauses, complement clauses, relative clauses) and do not take into account the different ways that language can be structurally complex (e.g., through phrasal embedding or the addition of non-finite clauses). This is especially important when it comes to comparing speech and writing. For example, finite verbal complements (e.g., "I think that…") tend to be more common in speech whereas relative clauses (e.g., "a ring which limits a central electrotransparent space") and other structures modifying noun phrases tend to be more common in writing (Biber et al., 2011: 20). In order to tap into cross-modal differences, it may therefore be more useful to look at phrase-level complexity features. Phrasal complexity may be more indicative of modal differences than clausal complexity because it allows for more dense packaging of information, which is hypothesized to be difficult in pressured online production. As summarized by Biber et al. (2021: 465):

> "There is a synergy between the production circumstances of the written mode (which enable a careful crafting of the text), the communicative purposes of informational academic discourse, and this ability to compress language with a high informational content into few words: taken together, these factors provide strong functional motivation for the dense use of phrasal complexity features."

In this instance Biber et al. were referring specifically to academic writing, but a similar argument could be made for all tasks that require communicating information in a succinct way. For instance, in the Fire Chief Task or vacation-planning tasks described in Section 2.3.2, learners need to consider many different aspects simultaneously which may require dense packaging of information and thus phrasal elaboration. This may explain why the studies that employed phrasal complexity measures (e.g., number of modifiers per noun phrase) did report significant differences between modes (cf. Vasylets et al., 2017).

For L2 learners, the developing state of the interlanguage system means that the formulator is under even more pressure to access items from the mental lexicon in speech as compared to writing, which may make dense phrasal elaboration even more difficult. In the case of L2 French, both De Clercq & Housen (2017) and Bartning et al. (2015) found that L2 speakers used, on average, shorter noun phrases than L1 speakers in an oral retelling task, suggesting that dense phrasal elaboration in the oral mode is indeed more difficult for L2 learners.[16]

If phrasal elaboration is in fact more challenging in the oral mode, it points towards a possible interaction effect between modality and task type on the one hand and proficiency on the other. This is exemplified by a study conducted by Biber, Gray and Staples (2016), who compared the syntactic complexity of written and oral TOEFL exam tasks. Within each modality, learners completed independent-type tasks, in which they had to build an argument based on their own ex-

[16] Of course, it is impossible to know whether phrasal elaboration is indeed more cognitively challenging on basis of corpus studies alone. As Biber et al. (2021) point out, there is a need to triangulate register-based corpus linguistic research with more experimental or psycholinguistic methods.

periences and integrated-type tasks, in which they had to synthesize information from a source passage. The results showed that, in line with the more personal or involved focus of these tasks, oral and independent tasks were associated with adverbs, adverbial clauses, desire verbs and coordinated clauses. Written and independent tasks, on the other hand, were associated with (among other things) the use of nouns, nominalizations, noun phrase modifiers and attributive adjectives, in line with both the informational nature of the tasks and the fact that writing seems to facilitate the online planning necessary for such dense information packaging. Moreover, the authors found a trend whereby higher scores for each component of the exam were associated with the use of more 'literate' linguistic features (as shown in Figure 2.6). This means that not only are phrasal complexity features associated with writing but because of the fact that they are associated with academic discourse, they also seem to be indicative of proficiency across the board in all types of tasks, even oral tasks (Biber et al., 2016).

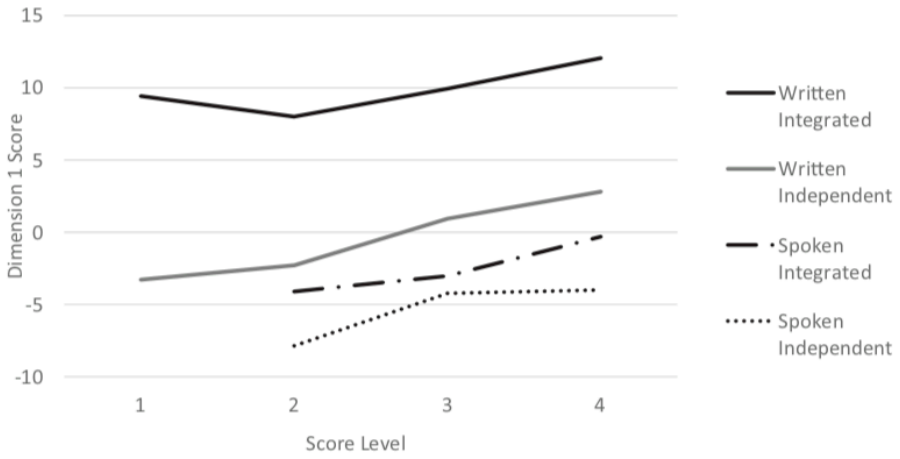


Figure 2.6: Dimension 1 scores (representing the use of oral vs. literate linguistic features) across task types and levels (Biber, Gray and Staples, 2016: 661)

Putting the results from the TBLT-based and MD-based research together, the picture that emerges is indeed that of an interaction between modality and register (or a "synergy" in the words of Biber et al.,

2021). In the case of lexical complexity, there is clear evidence across different tasks that writing seems to allow learners to ease burdens on the formulator thereby facilitating access to a more diverse and sophisticated set of lexical items. For syntactic complexity however, the communicative function of a task seems to be highly relevant to whether differences are observed between the two modes. Specifically, different tasks may promote the use of clausal versus phrasal complexity features. This is shown, for example, by a study conducted by Michel et al. (2019), who found that argumentative-type tasks exhibited, on average, more dependent clauses per T-unit compared to descriptive, instructional or narrative-type tasks. This would suggest that such tasks would have high levels of clausal complexity regardless of the modality they are completed in. For that reason, large-grained measures of subordination may not be that informative about cross-modal differences. On the other hand, if a task promotes the concise packaging of information, something that appears to be much easier in the written mode, then measures of complexity at the phrasal level may be more sensitive to cross-modal differences.

2.3.4 Modality and phraseological complexity

Thus far, the discussion of cross-modal differences has been focused on lexical and syntactic complexity but there is also reason to suggest that there may be phraseological differences between writing and speaking. For one, the fact that writing seems to reduce pressure on the formulator suggests that the written mode may promote higher levels of phraseological complexity overall because learners will be able to devote more attentional resources to linguistic output. There are clear parallels here with measures of lexical diversity and sophistication, which almost always tend to be higher in writing as compared to speaking (cf. Vasylets et al., 2017). When it comes to phraseology more specifically, the results of Forsberg and Fant (2010) and Erman et al. (2015) seem to indicate that the "difficulty" of a task may also play a role in how easily phraseological units can be retrieved, suggesting that performing a task through the written mode may help learners access phraseological units.

At the same time, studies of register variation demonstrate the importance of considering the communicative function of the task that learners are performing. Biber et al. (2004), for example, compared the number of lexical bundles between conversation, classroom teaching, textbooks and academic prose corpora. In addition to an overall higher quantity of lexical bundles in the oral corpora, consistent with other MD-based register research, the oral corpora were made up of dependent clause- and VP-based bundles, primarily used for expressing stance. In contrast, the written corpora were made up of NP-based referential bundles. In their study of TOEFL exam tasks, Staples et al. (2013) also found that different functional types of lexical bundles were used in oral as compared to written tasks. This means that if a learner does not use a diverse set of phraseological units, it is not necessarily because they are unaware of them, but rather, it could be because the task itself does not promote that type of unit. In fact, this was even noted by Durrant and Schmitt (2009) who found that the baseline quantity of premodifier-noun constructions differed across different sub-corpora, which the authors attributed to task differences. Forsberg and Fant (2010) also suggested that the reason that role plays exhibited fewer phrase-based lexical FS than narratives in their data was partially due to the pragmatic demands of that task (i.e., less of a need for specific "denotative-referential lexis"). In a similar way, a learner with 'unsophisticated' (low PMI) units may be responding to the demands of the task and purposefully selecting these units because they allow them to fulfill their communicative goals. One study that is illustrative in this respect is that of Paquot, Naets, et al. (2021), who found that essays on the topic of 'violent films' or 'lying,' contained verb + direct object dependencies with a higher PMI than essays about 'money,' which the authors attributed to the fact that the first two topics mobilized a larger semantic domain and thus allowed learners to draw on more sophisticated phraseological units (e.g., *commit + crime*, *serve + purpose*) than when they were writing about money (*have + money*, *buy + car*). While this study examined the effect of *topic* rather than task, it is reasonable assume that task type could also play a similar role in the mobilization of specific linguistic features, especially in light of the register variation literature showing that texts with more personal or involved in nature promote a more verbal discursive style

(Biber, 1988, 2014; Biber et al., 1999). This suggests that phraseological units that contain a verb are more likely to be attested in such tasks. In contrast, tasks that are more informational in purpose often require the dense packaging of information made possible through phrasal elaboration, which suggests that such tasks are more likely to be composed of phraseological units containing nouns. At the same time, the use of phrasal complexity features has also been shown to be associated with proficiency overall across both oral and written tasks because it is characteristic of academic discourse (Biber et al., 2011, 2016). This would indicate that noun-based phraseological units may be clearer indices of advanced development and proficiency in both written and oral tasks as compared to verb-based units. This goes to show that the relationship between modality and phraseological complexity on the one hand, and development/proficiency on the other, may not be entirely straightforward. In fact, as shown in the empirical studies that form the main chapters of this thesis, these relationships often tend to be, for lack of a better word, complex.

2.4 Conclusion

The preceding chapter has shown how Paquot's (2019) construct of *phraseological complexity* relates to the way that complexity has commonly been defined and operationalized in L2 research (Bulté & Housen, 2012). At the level of theory, the construct of phraseological complexity represents the intricate web of associations in the lexicon between lexical items and other linguistic elements (Gries, 2008), manifest through the observational constructs of diversity and sophistication and operationalized, for example, through (root) type-token ratios and the PMI of phraseological units that surface in learner production. Just as with lexical and syntactic complexity, these measures may reflect multiple constructs simultaneously and overlap with both absolute and relative forms of complexity.

A review of empirical evidence shows that phraseological complexity measures can be useful indices of proficiency, especially for highly advanced learners or when broad ranges of proficiency are compared cross-sectionally. The evidence for such measures as developmental

indices is, however, more mixed, with studies finding either no significant growth (Bestgen & Granger, 2014; Garner & Crossley, 2018; Kim et al., 2018; Yoon, 2016), very slow patterns of development (Edmonds & Gudmestad, 2021) or even a significant decrease in PMI over time (Siyanova-Chanturia & Spina, 2020). Likewise, the evidence for such measures as indices of oral proficiency is less convincing. Although some studies have found an association with oral proficiency (Eguchi & Kyle, 2020; Uchihara et al., 2021; Zhang et al., 2021), others have found PMI to be inversely correlated with proficiency (Paquot et al., in press; Tavakoli & Uchihara, 2020).

Just because there is no growth over time though, does mean that nothing is happening (Pallotti, 2009). Likewise, the fact that the strong relationship between phraseological complexity and proficiency found in studies of L2 writing do not always carry over to L2 speech does not mean that the measures themselves are not useful. Rather, as Pallotti points out, it is important to consider how these measures may be influenced by factors other than proficiency: "one should be aware that the fluctuations in CAF do not depend exclusively on psycholinguistic factors such as memory, automaticity of cognitive efficiency, but they may be responsive to the task's semantic and pragmatic demands" (Pallotti, 2009: 599). Staples et al. (2016: 177) make a similar argument in relation to syntactic complexity features across different academic disciplines: "the factors influencing the use of phrasal and clausal features are related not only to the level but also to the genre and discipline of these texts." This explains, for example, why commonly-used syntactic complexity measures often vary greatly between tasks (e.g., Alexopoulou et al., 2017; Michel et al., 2019). When it comes to considering the effect of modality, such differences are all the more relevant. While there seems to be clear evidence that the opportunity for greater online planning in the written mode can affect the ability for learners to select a more diverse and sophisticated set of words (and possibly also phraseological units), such an advantage is not a given for syntactic complexity because of the role of register. Syntactic features that help to meet a communicative goal (e.g., subordination in the case of argumentation) will need to be used by learners across modes if a task

requires it. Therefore in addition to cognitive factors, it is impossible to ignore the communicative function of a task when determining the possible effects of mode on phraseological complexity.

In the case of French more specifically, the preceding chapter showed how although the approach to phraseology differs in French, there are some important parallels with the findings for L2 English. In particular, the quantity and diversity of lexical FS has been shown to be correlated with proficiency in both writing and speaking (Erman et al., 2015; Forsberg, 2008, 2010; Forsberg Lundell & Lindqvist, 2014). Forsberg (2010) also showed that highly advanced learners and native speakers used lexical FS with a high PMI and the longitudinal results of Edmonds and Gudmestad (2021) showed that the PMI of collocations produced by university-level L2 writers increased slightly but significantly over a period of 21-months. Taken together, these results suggest that, just as in L2 English, phraseological complexity measures will be predictive of proficiency and development in L2 French.

That hypothesis is tested in all three main empirical studies of this thesis (Chapter 4, 5 and 6). However, just as in L2 English, modality is likely to play a role in the phraseological complexity of learner productions. This is shown, for example, by Granfeldt's (2007) study showing higher levels of lexical diversity in written as opposed to spoken L2 French. Importantly, Granfeldt's results also speak to the possible confound of register in the selection of syntactic structures and show that just as in L2 English, it will be important to separate out possible effects of register, given the functional and linguistic differences between spoken and written French (Blanche-Benveniste, 1995; Labbé, 2007). This is exactly what the empirical studies presented in Chapter 5 and Chapter 6 aim to do. When considered together, the studies in this thesis fill an important gap in the phraseological complexity literature and bring the issue of task and modality to the forefront in order to explore how they both contribute to the ability of phraseological complexity to index proficiency and development in L2 French.

Chapter 3

fsca: French syntactic complexity analyzer

Adapted from:

Vandeweerd, N. (2021). fsca: French syntactic complexity analyzer. *International Journal of Learner Corpus Research*, 7(2), 259–274.

Abstract

This chapter reports on an open-source R package for the extraction of syntactic units from dependency-parsed French texts. To evaluate the reliability of the package, syntactic units were extracted from a corpus of L2 French and were compared to units extracted manually from the same corpus. The F-score of the extracted units ranged from 0.53-0.97. Although units were not always identical between the two methods, manual and automatically-derived syntactic complexity measures were strongly and significantly correlated (ρ = 0.62-0.97, p < 0.001), suggesting that this package may be a suitable replacement for manual annotation in some cases where manual annotation is not possible but that care should be used in interpreting the measures based on these units.

3.1 Introduction

The development of syntactic complexity is a widely attested component of L2 proficiency (Ortega, 2003), the most common operationalizations of which include measures of length (e.g., mean length of clause), the ratio of one type of unit to another (e.g., clauses per sentence) and the diversity of syntactic units (e.g., standard deviation of clause length) (De Clercq & Housen, 2017; Wolfe-Quintero et al., 1998). In order to calculate these measures, texts must first be annotated for syntactic units at various levels of granularity (e.g., sentence, T-unit, clause, phrase), a process which is labour-intensive, especially considering the multiple annotators required to ensure reliability. Fortunately, for those researchers working on L2 English, automatic tools have been developed that can annotate and calculate syntactic complexity measures in learner texts. One such example is Lu's (2010) L2 Syntactic Complexity Analyzer (L2SCA), which uses the Stanford Parser (Klein & Manning, 2003) to segment, tokenize, part of speech (POS) tag and syntactically parse a text in terms of syntactic dependencies. A series of search expressions are then used to query the syntactic parse for syntactic units. Lu (2010) showed that there was a large degree of overlap between the manually and automatically identified units, with F-scores[1] ranging from 0.846 to 1.000. Moreover, there were strong correlations between syntactic complexity measures derived from the two identification methods. The use of such automatic tools is now common in studies of L2 English (e.g., Kyle & Crossley, 2018; Lu, 2011).

For languages other than English, the availability of such automatic tools is much more limited and texts must therefore be manually annotated for syntactic units. As shown in Table 3.1, this has meant that previous studies of L2 French have often needed to strike a balance between sample size (and text length) on the one hand and the number of syntactic units on the other (Benevento & Storch, 2011; Bernardini & Granfeldt, 2019; De Clercq & Housen, 2017; Gyllstad et al., 2014; Kuiken & Vedder, 2008; Way et al., 2000). This is unfortunate, given the impor-

[1] F-scores represent the balance between precision (not identifying too many incorrect units) and recall (not missing too many units that were identified by the manual annotation).

Table 3.1: Previous syntactic complexity studies in L2 French

	Number of texts					Manually annotated units						
Study	EN	IT	FR	Total	Length	AS	T	C	DP	CC	NP	IRR
Benvento and Storch (2011)	-	-	45	45	2-300*		x	x				100%
Bernardini and Granfeldt (2019)	20	20	20	60	n.r.		x	x	x			n.r.
De Clercq and Housen (2017)	100	-	100	200	n.r.	x		x	x	x	x	n.r.
Gyllstad et al. (2014)	108	56	76	240	n.r.		x	x	x			n.r.
Kuiken and Vedder (2008)	-	162	246	408	>150*		x	x	x			n.r.
Way et al. (2000)	-	-	937	937	47-118†		x					n.r.

Note:
EN = English, IT = Italian, FR = French; AS = AS-Unit, T = T-unit, C = Clause, DP = Dependent Clause, CC = Coordinated Clause, NP = Noun Phrase; IRR = interrater reliability of syntactic annotation; n.r.= not reported
* According to instructions in writing prompt
† The range of mean text lengths reported per group

tance of measuring complexity at various syntactic levels (see Norris & Ortega, 2009). What is also noticeable is that interrater reliability is only rarely reported and when it is reported (as in Benevento & Storch, 2011), very few details are provided about the process of double annotation. As such, there is currently no agreed-upon standard as to an acceptable level of reliability for syntactic annotation in L2 French.

In order to carry out the partial replication of Paquot (2018, 2019) presented in Chapter 4, it was necessary to calculate the same syntactic complexity measures as were used in the original study. Following the recommendation of Norris and Ortega (2009), we also wanted to investigate syntactic complexity at the T-unit, clause and phrase level but given the size of the corpus, manual annotation of those units was not feasible. Fortunately, because the learner texts had already been dependency parsed, it was possible to write a script in R (R Core Team, 2019) to extract syntactic units from the grammatical dependencies. The scripts were then compiled into a package (*fsca*) which is now available to download from GitHub.[2] This chapter describes the process that was used to evaluate the reliability of syntactic complexity measures based on the units that were extracted automatically with this tool.

[2] https://github.com/nvandeweerd/fsca

3.2 Methodology

In order to evaluate the automatic annotation tool, a set of learner data was first annotated for syntactic units (Section 3.2.1). These manual annotations served as a "gold-standard" against which the automatic extraction method was compared (Section 3.2.2). Reliability was evaluated by calculating the precision, recall and F-score for each syntactic unit (Section 3.3.1) and by correlating syntactic complexity measures calculated on the basis of manual versus automatically extracted units (Section 3.3.2).

3.2.1 Manual annotation

The source of learner data was the *Leerdercorpus Frans* (LCF, Demol & Hadermann, 2008; Vanderbauwhede, 2012), a corpus of texts written by university-level Dutch learners of French in Belgium. The subcorpus of argumentative essays was rated by two trained language assessors and all texts (n = 251) were found to range from level B2 to C2 of the Common European Framework of Reference (Council of Europe, 2001) (see details in Section 4.2.1). The original version of the corpus contains XML tags around each sentence. Three sets of segments were extracted on the basis of these tags: a set of 60 segments for training annotators, a set of 100 segments for testing interrater reliability and a final set of 400 segments for testing the automatic extraction. In order to evaluate on as diverse a sample as possible, these segments were randomly extracted from the argumentative subcorpus as a whole. The annotators were Master's students who are native speakers of French with experience in syntactic annotation from their coursework in linguistics and translation. They were trained to identify the following units: sentences[3], clauses, coordinated clauses, dependent clauses, T-units, noun phrases and verb phrases using a set of guidelines (see definitions in Appendix I).[4] Mistakes in the identifica-

[3] Following the extraction of the segments, it became apparent that the segments which had been pre-tagged as "sentences" in the original corpus sometimes contained more than one sentence according to our definition. Because of this, annotators were asked to manually identify sentences within each extracted segment.

[4] This is, of course, only a very restricted sample of grammatical structures and does not take into account the different functions that these structures may have (cf. Biber et al., 2020, 2021). The selection of units here was made in order to be able to replicate

Table 3.2: Interrater agreement for manual annotation of syntactic units in LCF data

Unit	Total	Agree	Disagree	%	π
Sentences	102	102	0	1.00	1.00
Clauses	203	194	9	0.96	0.94
Dependent Clauses	65	54	11	0.83	0.77
Coordinated Clauses	38	26	12	0.68	0.57
T-units	139	128	11	0.92	0.89
Noun Phrases	458	417	41	0.91	0.88
Verb Phrases	278	256	22	0.92	0.89

Note:
% = Percent agreement; π = Scott's (1955) correction for chance agreement

tion of units were discussed with the lead investigator and the guidelines were further refined during these meetings. After training, a second set of 100 sentences was used to test the interrater reliability of the manual annotation. Table 3.2 lists the percentage agreement for each syntactic unit as well as Scott's (1955) π, which applies a correction for chance agreement. The interrater agreement (π) for all units except dependent clauses and coordinated clauses was found to be above the minimally acceptable estimate of reliability for second language research (0.83) as recommended by Plonsky and Derrick (2016). The lowest levels of reliability were found for dependent clauses (π = 0.77) and coordinated clauses (π = 0.57). In most cases, disagreements on these units were due to inconsistent annotation (e.g., annotating only one of two coordinated clauses in a sentence or annotating the matrix clause separately from the dependent clause embedded within) rather than lack of knowledge about the segments themselves. Following the calculation of interrater reliability, all cases of disagreement were discussed and resolved and the guidelines were further refined, with a particular focus on dependent and coordinated clauses as they had been the largest source of disagreement.

The annotators then worked separately to annotate the final set of 400

the methodology of Paquot (2019) and therefore uses the same structures as that study despite an awareness of the limitations of the omnibus approach to syntactic complexity.

sentences. All cases of disagreement were resolved in meetings with the lead investigator, who also checked the final list for errors. The combined set of 500 segments (interrater reliability set plus the 400 set) was then used as a gold standard to evaluate the R package described in Section 3.2.2. All annotations were carried out using Dexter Coder (Garretson, 2011), a program in which texts are annotated by highlighting segments in different colors. The highlighted segments were then extracted in the form of an XML file for comparison to the automatically extracted units.

3.2.2 Automatic extraction of syntactic units

In the context of the study described in Chapter 4, the learner texts were POS tagged with MElt Tagger (Denis & Sagot, 2012) and parsed with Malt Parser (Nivre et al., 2006) trained on the French Tree Bank (Abeillé & Barrier, 2004), a parsing method which has been shown to have a high level of accuracy (87% labeled attachment) on L1 data (Candito, Nivre, et al., 2010). The choice of these tools was necessitated by the requirement to use the same processing chain as was used for the FRCOW16 reference corpus. The output of Malt Parser is a text file containing a list of tables, each corresponding to one sentence. Each table lists the lemma and part of speech for each word (token) in the sentence along with the dependency relationships between the words. This is illustrated in Table 3.3. In this case, the main verb of the sentence is *est* ('be-3SG.PRES') and is labeled as the 'root' of the sentence in the column *Dependency Relation*. As the top level node of a sentence, it is not dependent on any other word, hence the value of 0 in the *Dependent On* column. Because *est* is the second word of the sentence, the position is 2 (indicated in the *Position* column). All words which are directly dependent on *est* have the value of 2 in the *Dependent On* column. This includes the subject *C'* ('it') as well as the subject attribute *point* ('point'). The sentence-final period is also dependent on the root. The words that are directly dependent on *point* include the determiner *un* ('a') and the adjective *important* ('important'), which is itself modified by the adverb *très* ('very'). In this way, every word in the sentence is dependent on one and only one word.[5] Following parsing,

[5] For an overview of dependency parsing see Green (2011).

Table 3.3: Example dependency-parsed sentence

Token	Lemma	Part of Speech	Position	Dependency Relation	Dependent On
C'	ce	PRO:DEM	1	suj	2
est	être	VER:pres	2	root	0
un	un	DET:ART	3	det	4
point	point	NOM	4	ats	2
très	très	ADV	5	advmod_ADJ	6
important	important	ADJ	6	amod	4
.	.	SENT	7	ponct	2

each sentence was aligned with the manually identified segments by assigning the corresponding code from the XML manual annotation file.

The `getUnits()` function from the *fsca* package extracts syntactic units by first retrieving a list of the relevant node words for a given unit (e.g., nouns for noun phrases) and then extracting all of the dependencies on each of the node words using the `induced.subgraph()` function from the *igraph* package (Csardi & Nepusz, 2006). Additional cleaning steps are then performed (e.g., ensuring there is a finite verb, removing punctuation etc.) The output of the function is a list of syntactic units including the number of units, the length of each unit and the tokens belonging to each unit (see Example 3.1). The following syntactic units are extracted by the function: sentences, clauses, dependent clauses, coordinated clauses, T-units, noun phrases and verb phrases. More detailed information about the extraction of each of these units is provided in Appendix I.

Example 3.1. Example *fsca* output:

Ils prétendent qu'il est impossible de rééduquer un tel jeune criminel.

```
getUnits(test.sents[["b.208.1"]],
         units = c("DEP_CLAUSES"),
         paste.tokens = TRUE)
```

```
## $NUMBER
## [1] 1
##
## $LENGTHS
## [1] 10
```

```
##
## $TOKENS
## $TOKENS[[1]]
## [1] "qu' il est impossible de rééduquer un tel jeune criminel"
```

3.3 Results

3.3.1 Precision and recall of automatically identified units

To evaluate the reliability of the automatic method, precision, recall and F-scores for each syntactic unit were calculated according to the formulas used in Lu (2010: 486):

$$\text{precision} = \frac{\text{number of identical manual and automatic segments}}{\text{number of automatic segments}}$$

$$\text{recall} = \frac{\text{number of identical manual and automatic segments}}{\text{number of manual segments}}$$

$$\text{F-score} = \frac{2 * (\text{precision} * \text{recall})}{\text{precision} + \text{recall}}$$

As shown in Table 3.4, although the number of units identified by each method was quite similar, the number of identical units identified by the two methods varied across the unit types. Lower F-scores were found for dependent clauses (0.60) and coordinated clauses (0.53) and higher F-scores were found for clauses (0.74), T-units (0.83), sentences (0.97), noun phrases (0.84) and verb phrases (0.78).

Table 3.4: Precision and recall of automatically extracted units from LCF data

Unit	Manual	Automatic	Identical	Precision	Recall	F-score
Sentences	517	519	500	0.96	0.97	0.97
Clauses	903	871	652	0.75	0.72	0.74
Dependent Clauses	325	295	187	0.63	0.58	0.60
Coordinated Clauses	153	144	78	0.54	0.51	0.53
T-units	581	576	478	0.83	0.82	0.83
Noun Phrases	2277	2278	1903	0.84	0.84	0.84
Verb Phrases	1328	1317	1030	0.78	0.78	0.78

3.3.2 Correlation between manual and automatic methods

Eleven syntactic complexity measures were computed on each of these segments using both the manual and automatically identified units. These measures target complexity at the sentence, T-unit, clause and phrase-level in terms of mean length, number of embedded units and standard deviation (see Chapter 4). Pearson's r was calculated for all normally distributed measures (tested using Shapiro Wilks tests) and Spearman's rho (ρ) was calculated for all non-normally distributed measures. These are provided in Table 3.5. All correlations were found to be significant ($p < 0.001$) and are considered large according to Plonsky and Oswald's (2014) recommended guidelines for L2 research.

3.3.3 Sources of error

Each case of disagreement between the manual and automatic methods was annotated as being due to: incorrect segmentation (Seg.), an erroneous part of speech tag (POS), an incorrect dependency label or incorrect dependency relation (Dep.), a learner error (Learner) or an error due to scripts in the package (Fun.). The tabulation of these errors is provided in Table 3.6

As shown in Table 3.6, the majority of the errors occurred during preprocessing (segmentation, POS tagging and dependency parsing). At the segmentation stage, some sentences were split at non-sentence boundaries (e.g., semi-colons, ellipses) which caused otherwise con-

Table 3.5: Correlation between manual- and automatic-based (fsca) measures

Measure	Measure Description	ρ	r	p
MLS	Mean length of sentence	0.972	-	***
DIVS	Standard deviation of sentence length	-	0.961	***
T_S	T-units per sentence	0.645	-	***
MLT	Mean length of T-unit	0.871	-	***
DIVT	Standard deviation of T-unit length	0.620	-	***
C_T	Clauses per T-unit	0.823	-	***
MLC	Mean length of clause	0.887	-	***
DIVC	Standard deviation of clause length	0.759	-	***
MLNP	Mean length of noun phrase	0.720	-	***
DIVNP	Standard deviation of noun phrase length	0.713	-	***
NP_C	Noun phrases per clause	0.892	-	***

Note:
*** p < 0.001

Table 3.6: Errors in fsca annotation

	Preprocessing					
Unit	Seg.	POS	Dep.	Learner	Fun.	Total
Sentences	13	-	-	3	-	16
Dependent Clauses	1	3	81	9	52	146
Coordinated Clauses	5	1	31	3	21	61
Noun Phrases	1	23	331	12	74	441
Verb Phrases	1	9	143	11	98	262
Total	21	36	586	38	245	926

Note:
Errors in T-units and clauses were not annotated separately as they necessarily include the errors in the embedded units.
Seg. = Segmentation error; POS = Part of speech error; Dep. = Dependency parsing error; Learner = Learner error; Fun. = Function error

tiguous segments to be analyzed separately. POS-tagging errors also caused problems because the root nodes, which form the basis of the identification of syntactic units, could not be found. For example, mis-tagging the noun *jeune* ('young/teenager') as an adjective, meant that noun phrases that were dependent on *jeune* were not identified. This also caused problems for other units in which the noun phrase was embedded. The most common type of error was due to incorrect assignment of dependency relations (e.g., prepositional phrases that were analyzed by the parser as being dependent on the immediately preceding noun instead of the verb that they modify). This often caused a cascade of problems which meant that several units were misidentified or unidentified.

In addition, although the learner texts in the corpus are fairly advanced (B2-C2), there were still cases where learner errors caused problems for the parser. These included: punctuation errors (e.g., lack of space after a period), spelling errors (e.g., *son* ('his/her') for *sont*, 'be-3PL.PRES') and the omission of obligatory elements (e.g., lack of a finite verb in a dependent clause). Each of these errors caused POS-tagging and parsing errors which led to the misidentification of syntactic units by the function.

About a quarter of the errors were due to the current limitations of the function itself. Some particular structures that cause issues for the function include: verb phrases with adverbs between the auxiliary verb and the main verb; coordinated phrases; dependent clauses directly dependent on prepositions or adverbial interrogatives; non-nouns preceded by a determiner including superlatives and the non-functional determiner *l'* in *que l'on*; coordinated clauses with a pronoun as the subject of the verb; noun phrases modified by both a prepositional phrase and a relative clause; imperative clauses; exclamatory clauses headed by a subordinating conjunction and inverted comparatives (e.g., *Aussi divergeant que les affirmations sont les approches possibles.*, 'Just as diverging as the claims are the different approaches'). It is important to note however that not all cases of disagreement between the manual and automatic methods are necessarily equally problematic. For example, Malt Parser analyzes both words in multi-word conjunctions (e.g., *tandis que*; 'while') as modi-

fiers of the main verb and therefore only *que* is captured as part of the dependent clause. As a result, multi-word conjunctions are often analysed by the function as belonging to the main (or in the case of coordinated clauses, left) clause by the function instead of the dependent (or right) clause. In these cases, the only difference between the units identified by the two methods is the location of the conjunction. This may explain why complexity measures such as the mean length of clause are strongly correlated between the manual and automatic methods despite the low F-scores found for coordinated and dependent clauses.

3.4 Discussion and conclusion

The aim of this chapter was to determine the reliability of syntactic complexity measures calculated on the basis of syntactic units extracted using the *fsca* package. To answer this question, a sample of segments was extracted from a learner corpus of French and was manually annotated for the presence of six syntactic units. The manually annotated units were compared to a list of automatically extracted units. The F-score, which represents the balance between precision and recall, was found to range from 0.53-0.97 and correlations between manually- and automatically-based syntactic complexity measures were found to range from 0.62 to 0.97. In other words, the units extracted by the *fsca* package did not always match up identically with the units identified manually but the measures calculated on the basis of the units were still strongly correlated with manually-based measures. When compared to Lu's (2010) tool for L2 English (L2SCA), the reliability reported here is noticeably lower (Lu reports F-scores of 0.83-1.00 and correlations of 0.83-0.94). The question is whether this tool can be considered reliable, given the strength of these correlations.

While standard thresholds have been suggested for other types of reliability such as inter-item, interrater and intrarater (Brown, 2014; see Landis & Koch, 1977; Plonsky & Derrick, 2016; Shrout, 1998), there is currently no agreed-upon threshold for the reliability of automatic annotation in the field of learner corpus research. Among studies of syn-

tactic complexity in L2 French more specifically, only one study has reported a coefficient of reliability (and only interrater reliability) for this type of annotation. Benevento and Storch (2011), report 100% interrater agreement on the annotation of T-units and clauses, two units that were also found to have a high level of interrater reliability (π=0.89; π=0.94) as well as high F-scores in the automatic extraction (0.83 and 0.74) in this study. However, the authors do not report how many texts were double annotated and so it is difficult to directly compare these results. In addition, whether such high levels of agreement would also be found for other units such as dependent and coordinated clauses is unknown but perhaps doubtful given the low level of agreement obtained by the annotators in this study. It is also important to note that manual annotation is not necessarily unproblematic either. The results presented here suggest that even high levels of interrater reliability may be difficult to achieve for some types of syntactic units (especially dependent and coordinated clauses). This is especially troubling considering the fact that few studies report coefficients of interrater reliability for syntactic annotation.

In the absence of general thresholds of reliability for automatic annotation or comparable studies of syntactic annotation in L2 French, it may be more useful to consider each of the measures as containing more or less noise due to machine-induced error. Whether to use such measures ultimately depends on the amount of error that a researcher is willing to accept. In the current study, measures with the highest correlations (above .90) such as MLS and DIVS involve less noise while measures with lower correlations such as DIVT (ρ = 0.620) and T_S (ρ = 0.645) involve more noise and therefore more caution should be used when interpreting results based on these measures (see further discussion of this issue in Section 7.3.1).

In order to reduce such measurement error, it is vital that we continue to refine the accuracy of natural language processing tools used in Learner Corpus Research. This study serves as a reminder that there is an element of error at each step of processing a learner text. While individual errors may not be detrimental on their own, as our processing techniques become more advanced, we need to ensure more than ever that the tools we use are as accurate as possible (see the

recent special issue in the *International Journal of Learner Corpus Research* on this topic). This study revealed that even commonly used techniques such as sentence segmentation and POS tagging still require fine-tuning. It may be beneficial therefore to test the reliability of newer, state of the art natural language processing tools (e.g., spaCy 2, Honnibal & Montani, 2017) on learner data. However, as noted by an anonymous reviewer, it is also important to remember that these tools need to be tested on a variety of situations (e.g., L1/L2 data, corrected/uncorrected texts, lower/higher proficiency learners etc.) in order to obtain a clear picture of their accuracy for a given purpose. While that is beyond the scope of this chapter, this recommendation would undoubtedly serve the field. That being said, automatic annotation tools for languages other than English are sorely needed and this open-source package is shared with the hope that it will be further refined through collaboration between researchers willing to broaden the scope of L2 French learner corpus research beyond what is possible with manual annotation alone.

Chapter 4

Cross-sectional comparison of phraseological complexity in written production

Adapted from:

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(2), 197–229.

Abstract

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 was operationalized as the diversity (RTTR) and sophistication (PMI) of three types of phraseological units: adjectival modifiers, adverbial modifiers and direct objects. Results revealed 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 included 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.

4.1 Introduction

It is well-recognized that complexity plays an important role in the development of L2 proficiency along with accuracy and fluency (Housen et al., 2012; Housen & Kuiken, 2009; Skehan, 1996, 2009b). 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 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 as compared to texts written by L2 learners of English, suggesting that native writers were more sensitive highly exclusive 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 PMI 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 (CEFR, 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 PMI was a significant predictor of the proficiency level of the learner texts in their corpus.

According to Paquot (2019), PMI represents the sophistication of

knowledge that a learner has about word combinations. In addition to sophistication, 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 (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 significant differences between B2-, C1- and C2-level essays for phraseological sophistication measures but while some linear patterns were found for lexical and syntactic measures, none of these were found to be significant. A follow-up study also showed that the mean PMI of direct objects and adjectival modifiers together explained 25% of the variance in holistic ratings of the essays (Paquot, 2018)[1]. The diversity of the phraseological units was not found to be predictive of the CEFR level in either study.

Until now, phraseological complexity research has focused on L2 English (but for L2 Dutch see Rubin et al., 2021) 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 as-

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

sessments 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 chapter is to determine how phraseological complexity develops in a more synthetic language by partially replicating the results of Paquot (2019) on L2 English on a comparable sample of L2 learners of French.

4.1.1 Complexity research in L2 French

Complexity in traditional linguistic domains has shown to be associated with increased proficiency in L2 French. For example, 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 et al., 2006). Similar relationships with proficiency have been found for syntactic complexity in terms of length of syntactic units (De Clercq & Housen, 2017; Gyllstad et al., 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 (see also Bartning et al., 2015). 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 et al., 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. In their study of long-residency learners, Erman et al. (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 '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?

4.2 Methodology

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 (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

(2019) and the current study are highlighted in the following sections.

4.2.1 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* (LCF, Demol & Hadermann, 2008; Vanderbauwhede, 2012). Because complexity features can vary between tasks and registers of academic writing (e.g., Alexopoulou et al., 2017; Michel et al., 2019; Staples et al., 2016), the comparison of the results to those of Paquot (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 full corpus contains 253 argumentative essays written on seven different prompts (see Table 4.1)[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 (2019) which were 3000-3500 words on average.

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

[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.

Table 4.1: Argumentative writing prompts in LCF corpus

	Writing Topics
1.	La loi sur l'euthanasie doit-elle également s'appliquer aux mineurs (d'age)? *Should the euthanasia law also apply to minors?*
2.	Faut-il maintenir les centrales nucléares au-délà 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 come une justice sans juges (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?*

Table 4.2: Assessed proficiency levels in LCF corpus

			Text length	
Level	n	Tokens	Mean	SD
B2	26	13348	513.38	155.30
C1	106	51981	490.39	164.44
C2	37	19559	528.62	135.06

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

but only the global proficiency score was used for the subsequent analysis following Paquot (2019). The raters reached a high level of agreement on this score ($\kappa = 0.837$, $p < 0.001$). Texts that 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 4.2 lists the number of texts, total number of tokens and average text length at each level within the final corpus.

Table 4.3: Precision and recall scores for dependencies extracted from LCF corpus

Dependency	Precision	Recall	F-score
Adjectival modifier	0.90	0.73	0.81
Adverbial modifier	0.73	0.88	0.80
Direct object	0.90	0.75	0.82

4.2.2 Complexity measures

Phraseological complexity. As in Paquot (2019), phraseological complexity was operationalized as the diversity and sophistication of three types of dependency relations: adjectival modifiers (AMOD; adjective + noun), adverbial modifiers (ADVMOD_V; adverb + verb) and direct objects (DOBJ; verb + direct object). In order to extract the dependencies, the learner texts were POS-tagged with MElt POS Tagger (Denis & Sagot, 2012) and dependency parsed using Malt Parser (Candito, Crabbé, et al., 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 F-Score of 0.80 or greater for the automated annotation of each dependency relation of interest (see Table 4.3).

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 refer-

Table 4.4: Phraseological complexity measures

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	∑ PMIamod / Namod
Mean PMI for ADVMOD_V dependencies	∑ PMIadvmod/ Nadvmod
Mean PMI for DOBJ dependencies	∑ PMIdobj / Ndobj
Sophistication: Collocation Bands	
Proportion of non-coll AMOD dependencies	∑ (PMIamod <3) / Namod
Proportion of low-coll AMOD dependencies	∑ (PMIamod 3 ≤ 5) / Namod
Proportion of med-coll AMOD dependencies	∑ (PMIamod 5 ≤ 7) / Namod
Proportion of high-coll AMOD dependencies	∑ (PMIamod 5 > 7) / Namod
Proportion of non-coll ADVMOD_V dependencies	∑ (PMIadvmod <3) / N advmod
Proportion of low-coll ADVMOD_V dependencies	∑ (PMIadvmod 3 ≤ 5) / Nadvmod
Proportion of med-coll ADVMOD_V dependencies	∑ (PMIadvmod 5 ≤ 7) / Nadvmod
Proportion of high-coll ADVMOD_V dependencies	∑ (PMIadvmod 5 > 7) / Nadvmod
Proportion of non-coll DOBJ dependencies	∑ (PMIdobj <3) / Ndobj
Proportion of low-coll DOBJ dependencies	∑ (PMIdobj 3 ≤ 5) / Ndobj
Proportion of med-coll DOBJ dependencies	∑ (PMIdobj 5 ≤ 7) / Ndobj
Proportion of high-coll DOBJ dependencies	∑ (PMIdobj 5 > 7) / Ndobj

Note:
T = type; N = token; PMI = Pointwise Mutual Information; AMOD = Adjectival modifier; ADVMOD_V = Adverbial modifier; DOBJ = Direct object

ence 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 that occurred more than five times in the reference corpus were included in the calculation. Dependency pairs that 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 ≤ 5), medium (MI 5 ≤ 7), and high (MI > 7). Table 4.4 lists the phraseological measures.

Lexical complexity. In order to maximize comparability with Paquot (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 (2019), we calculated the diversity of all lexical (i.e., content) 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 (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 as well as in-house R scripts.

Further following Paquot (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. Frequency lists based on this corpus have previously been shown to discriminate between proficiency levels both in receptive vocabulary tests (Batista & Horst, 2016; Peters et al., 2019) as well as in oral production (e.g., De Clercq, 2015). As in Paquot (2019), the FFD-based measures were corrected for text-length by applying Guiraud's correction. 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

Table 4.5: Lexical complexity measures

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	$T_{\text{FFD-offlist}}$lex/√Nlex
Proportion of FFD off-list verb lemmas	$T_{\text{FFD-offlist}}$verb/√Nverb
Proportion of FFD off-list noun lemmas	$T_{\text{FFD-offlist}}$noun/√Nnoun
Proportion of FFD off-list adjective lemmas	$T_{\text{FFD-offlist}}$adj/√Nadj
Proportion of FFD off-list adverb lemmas	$T_{\text{FFD-offlist}}$adv/√Nadv
Proportion of LT on-list verb lemmas	$N_{\text{LT-onlist}}$verb/Nverb
Proportion of LT on-list noun lemmas	$N_{\text{LT-onlist}}$noun/Nnoun
Proportion of LT on-list adjective lemmas	$N_{\text{LT-onlist}}$adj/Nadj

Note:
T = type; N = token; TTR = Type-token ratio; 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)

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 that 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 4.5 lists the lexical complexity measures.

Table 4.6: Syntactic complexity measures

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

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).[3] Because no tool exists to calculate these measures automatically for French and because manually annotating all the corpus texts was not feasible, we created the *fsca* package to automatically annotate the corpus texts for syntactic units (Chapter 3). Table 4.6 lists the syntactic complexity measures.

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

[3]We are aware of the limitations of an 'omnibus' approach to syntactic complexity (see Biber et al., 2020, 2021) but as the principle goal in using such measures here was to compare the broad contribution of different domains of complexity (lexical, syntactic, morphological, phraseological) to proficiency scores, we feel that the use of such measures is merited (and follows the use of such measures in the study we are replicating).

(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 4.1.1). The calculation was based on the inflectional information provided by MElt POS tagger (Denis & Sagot, 2012) which tags each verb 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 morpho-semantic 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 (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."

4.2.3 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).[4] 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 non-collocation band) or "NOT_ALL" (indicating that at least one adverbial modifier 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. χ^2 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 χ^2 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 (α = 0.05/5 = 0.01). For the remaining non-binary phraseological variables, the Shapiro-Wilk test of normality and Levene's test of homogeneity of variance revealed that the assumptions for parametric tests could not be met so Kruskal-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 $r = \frac{z}{\sqrt{n}}$ (Rosenthal, 1994). Bonferroni corrections were again used to correct for multiple comparisons. To compare the three mean-PMI based measures, α was set at 0.017 (0.05/3). Likewise, α 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 (i.e., topic) 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

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

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 II. 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 et al., 2006; Strobl et al., 2007, 2008). We were primarily interested in comparing the relative contribution of different types of measures (i.e., phraseological, lexical, syntactic and morphological), rather than obtaining a predictive model that would generalize to other data sets. For this reason, we decided not to use a train-test split but to use all of the data to build the model. According to the authors of the *party* package, the statistical approach employed by the random forest model also ensures that "no form of pruning or cross-validation whatsoever is needed" (Hothorn et al., 2022: 9), so we have not done so here. The results reported below therefore relate to the classification accuracy of the model for the 78 sampled LCF texts only (i.e., not on a set of withheld texts). That being said, we also checked the classification accuracy of the model on two alternate random samples of LCF texts to ensure that the model would generalize over the whole data set. To determine which variables contributed to the classification of the texts, we used the *vip* package (Greenwell et al., 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 the *pdp* package (Greenwell, 2017) for each of these measures.

4.3 Results

4.3.1 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, which relate to linguistic and legal topics (Example 4.1). The high PMI dependencies also tended to be composed of words with very few other collocates (e.g., atténuant; 'attenuating').

Example 4.1. Units related to the topic of the essays:

- *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 Example 4.2:

Example 4.2. Units related to idioms:

- *bouillir+marmite* ('boil+pot'; = lit. to maintain a household)
- *promettre+monts* ('promise+mountains'; = lit. 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 FRCOW). In contrast, the dependencies with lower PMIs (Example 4.3) tended to belong to more general semantic domains and to be composed of higher frequency words such as *être* (be.INF; 11261/million words in FRCOW). 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').

Example 4.3. Examples of units with PMI less than 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 Table 4.7.

As shown in Table 4.8 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.

As shown in Table 4.9, 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

Table 4.7: Example dependencies extracted from LCF corpus

Example dependencies (PMI)
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)
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)
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)

Note:
ADVMOD_V = Adverbial modifier; AMOD = Adjectival modifier; DOBJ = Direct object

Table 4.8: Binary phraseological measures in LCF data (α = 0.01)

Unit	Type	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	-	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

Note:
ADVMOD_V = Adverbial modifier; AMOD = Adjectival modifier; DOBJ = Direct object

Table 4.9: Non-binary phraseological measures in LCF data

	Mean (SD)					
Measure	B2	C1	C2	χ^2	α	p
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
DOBJ_RTTR	4.3(0.81)	4.18(0.9)	4.53(0.79)	3.9	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	0.03
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
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

Note:
ADVMOD_V = Adverbial modifier; AMOD = Adjectival modifier; DOBJ = Direct object; RTTR = Root type-token ratio; PMI = Pointwise Mutual Information
* Significant for C1-C2(r=.21)
† Significant for B2-C1(r=.13) and B2-C2(r=.33)

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).

4.3.2 Random forest model (RQ2)

The random forest model had a classification accuracy of 0.96 which was significantly better ($p < 0.001$) than baseline. As shown by the confusion matrix in Table 4.10, the model was able to correctly classify most texts. All mis-classified texts were at most one level away from correct classification. On the two alternate samples of LCF texts,

Table 4.10: Confusion matrix for random forest model

	Reference		
Prediction	B2	C1	C2
B2	26	1	0
C1	0	24	1
C2	0	1	25

Note:
Note that this table reports the accuracy of the model on the sampled LCF texts (n = 78) and not on a separate set of (unseen) data.

the model had a classification accuracy of 0.78 and 0.86 respectively, indicating that the model could also generalize reasonably well over the whole data set.

The lowest variable importance score was -0.00296, so only variables with a VI higher than +0.00296 were considered important in the model (Levshina, 2015). As shown in Table 4.11, 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 4.1. 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 red line = B2, dashed blue line = C1, solid green 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 that is most probable at each value of the complexity measure. These ranges are also provided in Table 4.11. 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 use a greater variety of verbal inflections 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.

4.4 Discussion

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

Phraseological diversity was operationalized as the RTTR 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

Table 4.11: CEFR level prediction ranges for LCF texts according to partial dependency plots

		Prediction Ranges		
Variable	VI	B2	C1	C2
VAR_NOUN	0.00854	0.27-0.28	0.15-0.27	0.29-0.38
MCI.V	0.00750	2.74-5.46		5.58-8.66
FFD.RTTR.SOPH.LEX	0.00642	0.71-2.49	2.57-3.19	3.27-4.58
DOBJ_MEAN_PMI	0.00605	0.42-1.7	1.76-2.08	2.13-3.09
FFD.RTTR.SOPH.NOUN	0.00466	1.94-2.31	0.34-1.87, 2.38-2.53	2.6-3.99
VAR_LEX	0.00414	0.69-0.72	0.52-0.68	0.72-0.8
ADVMOD_V_PROP_PMI (NONCOL)	0.00392		0.97-1	0.87-0.97
MLNP	0.00366	3.01-3.37	3.41-5	

Note:
VAR_NOUN: Noun variation
(lexical diversity)
MCI.V: Morphological Complexity Index (for verbs)
(morphological diversity)
FFD.RTTR.SOPH.LEX: Proportion of FFD off-list lexical lemmas
(lexical sophistication)
DOBJ_MEAN_PMI: Mean PMI for DOBJ dependencies
(phraseological sophistication)
FFD.RTTR.SOPH.NOUN: Proportion of FFD off-list noun lemmas
(lexical sophistication)
VAR_LEX: Lexical word variation
(lexical diversity)
ADVMOD_V_PROP_PMI_NONCOL:
Proportion of non-collocating ADVMOD_V dependencies
(phraseological sophistication)
MLNP: Mean length of noun phrases
(syntactic complexity)

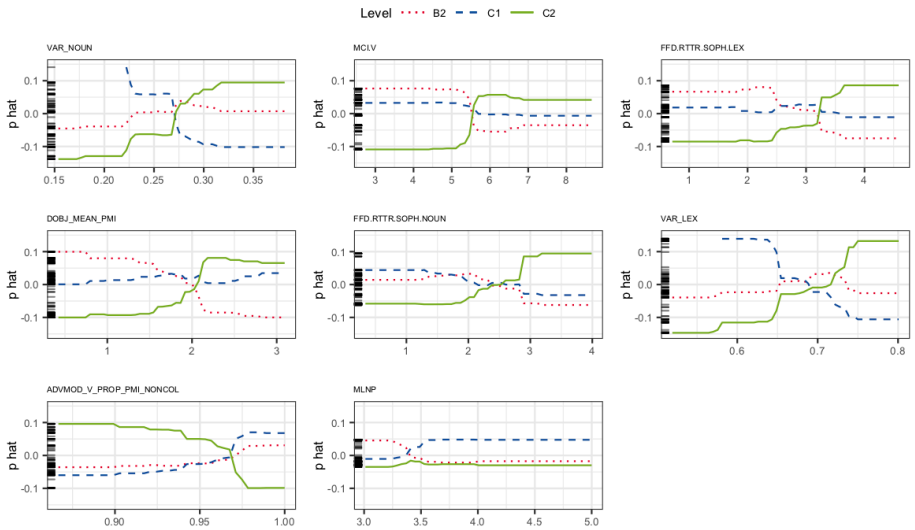


Figure 4.1: Partial dependency plots for important variables (VI > 0.00296). See following page for descriptions of measures.

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 4.4.2 we discuss this finding further in relation to another domain of complexity, namely morphological complexity.

Phraseological sophistication was operationalized as the mean PMI 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.

4.4.2 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 et al., 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 (2019) results for L2 English, whose final model contained only phraseological measures. Again, the differences between this study and Paquot (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 (2016) and Bartning et al. (2015) 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.[5] At first glance, it may be surprising that the diversity of 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

[5] A recent study of L2 Dutch found that the diversity of adverbial modifiers was a significant predictor of passing proficiency exams (Rubin et al., 2021). 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 significant in Paquot's (2019) study or in this study of L2 French, which both focused on learners beyond the B1 level.

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 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.

4.5 Conclusion

The main aim of this chapter 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 us-

ing 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 were also undertaken (Chapter 5 and 6 in this thesis). 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.

Chapter 5

Longitudinal development of phraseological complexity in oral and written production

Adapted from:

Vandeweerd, N., Housen, A., & Paquot, M. (in revisions). Comparing the longitudinal development of phraseological complexity across oral and written tasks. *Studies in Second Language Acquisition*.

Abstract

This study builds upon previous research investigating the construct validity of phraseological complexity as an index of L2 development and proficiency (Paquot, 2019). Whereas previous studies have focused on cross-sectional comparisons of written productions across proficiency levels, the current study compares *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 (cf. Housen & Kuiken, 2009). 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.

5.1 Introduction

The use of complexity measures as indices of L2 development has a long history in second language research (see for example 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: 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 of 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 lemma of a lexical item and one or more additional linguistic elements 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: 6). Such units have been shown to be an important component of the development of L2 proficiency. For example, Granger

and Bestgen (2014) showed that advanced learners used a higher proportion of collocations with a high pointwise mutual information (PMI)[1] as compared to beginner learners and Bestgen and Granger (2018) showed that the proportion of high PMI collocations increased in learner production over time.

While the authors of these studies did not use the term complexity themselves, Paquot (2019) argues that in measuring collocational strength in learner productions, 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 et al. (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

[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.

objects (e.g., win + lottery). Each text in the corpus was manually assessed by a minimum of two 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. We found similar results in a partial replication study using L2 French argumentative essays (B2-C2) (Vandeweerd et al., 2021, Chapter 4 of this thesis). 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) can be a useful index of L2 proficiency and development at the upper levels[2] of proficiency but they are limited by the fact that the focus has been exclusively on the written mode. In order 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 above 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 in order 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, Naets, et al., 2021; Qi & Ding, 2011), no study has yet directly compared the development of phraseology across oral and writ-

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

ten 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. (in press) 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 constrains 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 (Skehan, 2014). Along these lines, several studies have found higher levels of lexical diversity and sophistication in writing as compared to speaking in L2 production (e.g., 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

[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.

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 (N. C. 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 actually have similar linguistic features because they serve a similar communicative purpose (Biber, 2014). 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 & Conrad, 2019). Texts that are more informational in purpose on the other hand, 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 NP-based "referential bundles," which the authors attribute 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.

In regards to 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.

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: 204). However, it should be noted that the inclusion of native data is not meant to serve as a goal towards which the learners should strive (cf. recent criticisms of native-speakerism in L2 teaching and research by, among others, Holliday, 2006; 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: 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 actually required in order to complete a task successfully and to date, no study of phraseological complexity has included a native benchmark group.

[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.

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?

5.2 Methodology

5.2.1 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 learner group 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 participants is 21 (range 20-24) and includes 26 females and three males.[6] Prior to participation in the project, the learners had on average 10.5 years (range 9-15) of previous studies in French in an institutional setting. In addition to 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

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

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

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 for example 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 5.1). 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 below, each of which was repeated once during the study:

1. 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?
2. Pensez-vous que la marijuana devrait être légalisée?
   Do you think that marijuana should be legalized?
3. Pensez-vous que, de manière à inciter les gens à manger sainement, on devrait taxer les boissons sucrés 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?

They 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 designed to elicit perfective and habitual events. They were set in the past and contained a similar number of

Table 5.1: LANGSNAP data collection schedule (Adapted from Mitchell, Tracy-Ventura and McManus, 2017: 57)

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

colour images depicting two characters.

The oral interview was a semi-structured interview administered by one of the researchers. Each interview followed a list of pre-established questions about the participants' anticipations and experiences during their sojourn abroad. Questions were designed to elicit discussion of present, future, past and hypothetical events.

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 5.2 shows the median text lengths for each of the task types following preprocessing.

5.2.2 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 par-

Table 5.2: Text lengths across LANGSNAP tasks at each collection point

	Median (IQR)		
Cycle	Argumentative Essay	Oral Interview	Oral Narrative
Presojourn 1	210 (30)	1073 (475)	344 (120)
Insojourn 1	205 (24)	1630 (955)	409 (143.5)
Insojourn 2	209 (23)	1476 (888)	277 (112)
Insojourn 3	216.5 (32.25)	1105 (786.5)	324.5 (116.75)
Postsojourn 1	211 (30)	692 (264)	402 (69)
Postsojourn 2	217 (12)	812 (370)	253 (92)

ticular 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 which 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

Table 5.3: Precision and recall scores for units extracted from LANGSNAP corpus

	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

essays and 100 utterances[7] from the oral tasks was annotated by two researchers for the presence of adjectival modifier and direct object relations following the definitions in Appendix III. The annotators reached a high level of agreement for both units in both written (κ_{amod} = 0.92, κ_{dobj} = 0.84) and oral data (κ_{amod} = 0.94, κ_{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 5.3, 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.

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 5.2), 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 was done because

[7] The oral data was segmented into utterances following the CHILDES conventions (see MacWhinney, 2020). These utterances are henceforth referred to as *sentences*.

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

transformations of type-token ratios such as Guiraud's (1954) root type-token ratio (as used by Paquot, 2019; Vandeweerd et al., 2021) were found to be strongly 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. Because the absence of specific phraseological units (e.g., adjectival modifiers in oral productions) could be indicative of important register-induced variation (in line with Biber et al., 2004), we decided not to exclude these texts but to use a moving average method instead. This is similar to a commonly used method used to control for text-length differences when calculating lexical diversity (MATTR, Covington & McFall, 2010).[9] To maintain comparability between the diversity 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 score[10] 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 is FRCOW16 (Schäfer, 2015; Schäfer & Bildhauer, 2012): a ten billion word corpus of French. The advantage of this corpus is that it is provided with dependency annotation from which adjectival modifiers and direct objects can be automatically extracted using R scripts. In addition, it contains POS tags generated by Treetagger, the same POS tagger used to process the learner corpus. Following the method described by Paquot (2019),

[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 (τ < 0.01) between position in a text and the number of units, suggesting that the units were fairly evenly distributed throughout texts.

[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: 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.

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.

5.2.3 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 above: (1) adjectival modifier types, (2) adjectival modifier PMI, (3) direct object types and (4) direct object PMI. In each model, the phraseological complexity variable was the dependent variable. 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 included an interaction effect between EIT.PRE (initial proficiency) and time in months as a control variable. 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 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 modelling, the variables EIT.PRE (pre departure EIT scores) and COSINE.log (log of cosine similarity) were converted to z-scores. The independent variables in the model were time (in months), task

type and the interaction between time and task type. 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 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 which 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.

5.3 Results

5.3.1 The Learner Models

Adjectival modifier diversity. As shown in Table 5.4, 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 (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 Table 5.5, 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,

Table 5.4: Adjectival modifier diversity in LANGSNAP learner data

Fixed Effects	b	95% CI	SE	df	t	p
(Intercept)	2.162	[1.903,2.421]	0.132	477	16.35	<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,-0.57]	0.155	477	-5.638	<0.001
TASKTYPEnar→int	0.35	[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]	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				

Note:
COSINE.log = Logged cosine similarity with first argumentative essay
EIT.PRE = Elicited Imitation Score at presojourn
Call: AMOD.TYPES = COSINE.log + (EIT.PRE * MONTHS) + (TASKTYPE * MONTHS)
Marginal R^2:0.631
Conditional R^2:0.655

Table 5.5: Linear trends for adjectival modifier diversity in LANGSNAP 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]				

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.

Adjectival modifier sophistication. As shown in Table 5.6, the only significant predictor of adjectival modifier sophistication (PMI) was task type (written vs. oral). The average PMI of adjectival modifiers

Table 5.6: Adjectival modifier sophistication in LANGSNAP learner data

Fixed Effects	b	95% CI	SE	df	t	p
(Intercept)	1.097	[0.803,1.391]	0.15	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	[-0.116,0.3]	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	-0.01	[-0.03,0.009]	0.01	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				

Note:
COSINE.log = Logged cosine similarity with first argumentative essay
EIT.PRE = Elicited Imitation Score at presojourn
Call: AMOD.MEANPMI = COSINE.log + (EIT.PRE * MONTHS) + (TASKTYPE * MONTHS)
Marginal R^2:0.067
Conditional R^2:0.08

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

Direct object diversity. As shown in Table 5.7, 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 non-significant. This model explained 26.16% of the variance in direct object types.

Direct object sophistication. As shown in Table 5.8, significant predictors of direct object sophistication (PMI) included task type (narrative

Table 5.7: Direct object diversity in LANGSNAP learner data

Fixed Effects	b	95% CI	SE	df	t	p
(Intercept)	3.64	[3.333,3.948]	0.157	477	23.206	<0.001
COSINE.log	-0.03	[-0.115,0.054]	0.043	477	-0.709	0.479
EIT.PRE	-0.03	[-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]	0.08	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	-1.39	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				

Note:
COSINE.log = Logged cosine similarity with first argumentative essay
EIT.PRE = Elicited Imitation Score at presojourn
Call: DOBJ.TYPES = COSINE.log + (EIT.PRE * MONTHS) + (TASKTYPE * MONTHS)
Marginal R^2:0.201
Conditional R^2:0.262

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.

5.3.2 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 5.9, adjectival modifier diversity (number of types) was significantly higher in the argumentative essays (Mean=3.47, 95% CI=2.89,4.04) as compared

Table 5.8: Direct object sophistication in LANGSNAP learner data

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,-0.01]	0.024	477	-2.353	0.019
EIT.PRE	0.055	[-0.02,0.129]	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	[-0.21,-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	[-0.02,0.006]	0.007	477	-1.105	0.270
MONTHS:TASKTYPEnar→int	-0.004	[-0.011,0.003]	0.004	477	-1.05	0.294
Random Effects	**Variance**	**sd**				
(Intercept)	<0.001	0.021				
MONTHS	<0.001	0.004				
Residual	0.275	0.524				

Note:
COSINE.log = Logged cosine similarity with first argumentative essay
EIT.PRE = Elicited Imitation Score at presojourn
Call: DOBJ.MEANPMI = COSINE.log + (EIT.PRE * MONTHS) + (TASKTYPE * MONTHS)
Marginal R^2:0.136
Conditional R^2:0.144

to the narratives (Mean=1.12, 95% CI=0.78,1.46) and the interviews (Mean=1.44, 95% 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, 95% CI=1.8,2.5) and the narratives (Mean=1.93, 95% CI=1.54,2.33) than in the interviews (Mean=1.2, 95% 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, 95% CI=3.53,4.89) and the narratives (Mean=3.86, 95% CI=3.44,4.29) than in the interviews (Mean=2.98, 95% CI=2.76,3.2) 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, 95% CI=1.39,1.72) as compared to the interviews (Mean=1.09, 95% CI=0.83,1.35). The mean PMI of direct objects was also higher in the essays (Mean=1.54, 95% CI=1.19,1.9) 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

Table 5.9: Task differences in LANGSNAP native data

	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 (1.8,2.5)	1.93 (1.54,2.33)	1.2 (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 (2.76,3.2)	1 (0.578)	4.01 (0.001)	4.53 (<0.001)
DOBJ.MEANPMI	1.54 (1.19,1.9)	1.55 (1.39,1.72)	1.09 (0.83,1.35)	-0.07 (0.997)	2.33 (0.060)	3.44 (0.003)

Note:
AMOD = Adjectival modifer; DOBJ = Direct object

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.

5.4 Discussion

5.4.1 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 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 5.1, 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 5.2) or events occurring in their own lives in the case of the interviews (Example 5.3). 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).

Example 5.1. Argumentative Essay (G126d):

> "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.*

Example 5.2. Oral Narrative (B100c):

> "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.*

Example 5.3. Oral Interview (0118a):

> "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.*

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 written mode promotes adjectival modifiers in general and certain essay topics require 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 on the other hand, 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 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 (R. 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 favoured 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., in press) 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 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.

5.4.2 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 5.1. 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 grey shading). As shown by the figure, the two dimensions of phraseological complexity, namely diversity and sophistication showed different patterns of development. For diversity, 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 of time (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.

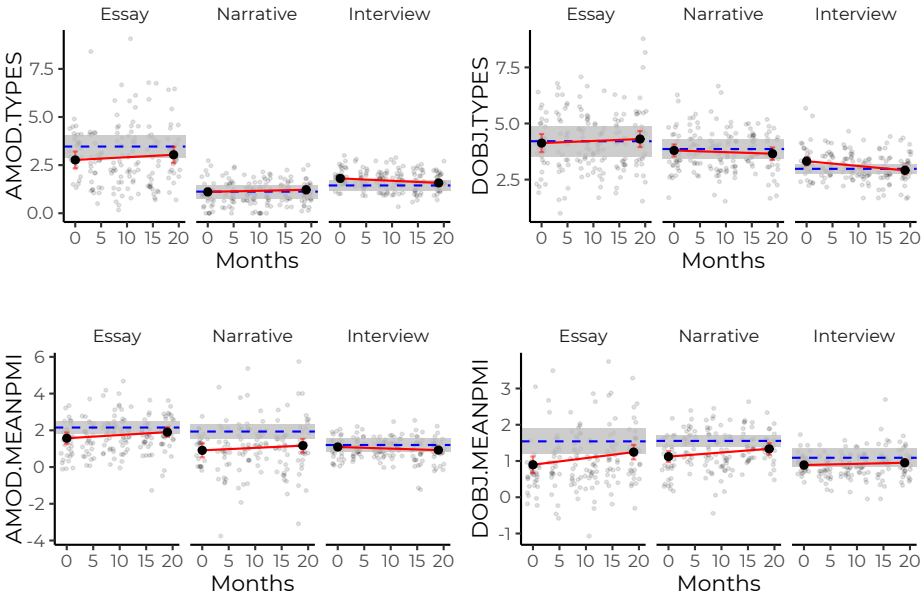


Figure 5.1: Predicted estimates and raw values of phraseological complexity measures over time (solid red line) compared to L1 benchmark (blue dashed line with 95 percent confidence intervals indicated by grey shading); AMOD = Adjectival modifier; DOBJ = Direct object

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 pre-departure and the second post-stay collection nineteen months later. Although the Edmonds and Gudmestad 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: 11) report a monthly increase of 0.04,[12] 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 of between the three tasks was not significantly different for the PMI of direct objects, as can be seen in Figure 5.1, the estimated marginal trend for both the essay (0.018) and the narrative (0.011) 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 in-

[12] Dividing the estimate from pre-departure to post-stay (0.7433) by the number of months between collection points (19)

terviews 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 predeparture 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* ('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 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 re-use units they are comfortable with. 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 it masks vocabulary gaps and artificially raises phraseological complexity at earlier time points (cf. Paquot et al., in press).

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 on the other hand, 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.

5.5 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 (R. 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 actually 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 differences in rate of development were 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 de-

velopment (De Bot & Larsen-Freeman, 2011). 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 (De Bot & Larsen-Freeman, 2011), require many more data collection points than the six that are available in the LANGSNAP data.

Another limitation is that in order 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, Naets, et al., 2021), we believe that the generalizability gained by including random slopes and random intercepts outweighs this downside. In an attempt to control the topic, we removed the phraseological units that were directly provided in the prompts or the picture stories and we included the cosine similarity between all texts and the argumentative essays as a control variable 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 modelling as proposed by Murakami et al., 2017).

Lastly, the scope of this study is limited in that we have chosen to focus on only two types of phraseological units. This is not meant to imply that the measures presented here capture the entire construct of phraseological complexity in French. Indeed, it is important that future studies begin to expand the types of phraseological units under investigation in order to achieve a more representative picture of a learner's phraseological repertoire. Chapter 6 of this thesis represents

an attempt to do just that.

As a way of concluding, it is interesting to relate these findings to Skehan's (2009a) discussion of the various factors that influence the *lexical* complexity of learner productions, namely: the development of vocabulary size and/or organization, performance conditions (e.g modality), style and other task influences (e.g., interactivity). This study has shown that in the case of *phraseological* complexity, similar factors appear to be at work. While the results of this study and others have shown that development and proficiency are important determinants of phraseological complexity, we have thus far only begun to scratch the surface in our understanding of how task-related factors also contribute to measures of phraseological complexity.

Chapter 6

Cross-sectional comparison of phraseological complexity in oral and written production

Adapted from:

Vandeweerd, N., Housen, A., & Paquot, M. (in revisions). Mes salutations les plus distinguées: What the lexis-grammar interface can reveal about proficiency on oral versus written French exam tasks. *Language Testing*.

Abstract

This study investigates whether re-thinking the separation of lexis and grammar in language testing could lead to more valid inferences about proficiency across modes. As argued by Romer (2017), typical scoring rubrics ignore important information about proficiency encoded at the lexis-grammar interface, in particular how the co-selection of lexical and grammatical features is mediated by communicative function. This is especially evident when assessing oral versus written exam tasks, where the modality of a task may intersect with register-induced variation in linguistic output (e.g., Biber et al., 2016). This article presents the results of an empirical study in which we measured the diversity and sophistication of four-word lexical bundles extracted from a corpus of French proficiency exams. Analysis revealed that the diversity of noun-based bundles was a significant predictor of written proficiency scores and the sophistication of verb-based bundles was a significant predictor of proficiency scores across both modes, suggesting that communicative function (Biber et al., 2004) as well as the constraints of online planning (Skehan, 1998) mediated the effect of lexicogrammatical phenomena on proficiency scores. Importantly, lexicogrammatical measures were better predictors of proficiency than solely lexical-based measures, which speaks to the potential utility of considering lexicogrammatical competence on scoring rubrics.

6.1 Introduction

Recently, there have been calls to re-think traditional divisions between "vocabulary" and "grammar" that have been typical in models of language ability used in language testing (e.g., Alderson & Kremmel, 2013; Kremmel et al., 2017; Römer, 2017). As argued by Römer (2017), language proficiency rating scales often fail to acknowledge the fact that lexis and grammar interact in meaningful ways. This is evident in corpus linguistic data showing that certain words and grammatical structures demonstrate patterns of co-occurrence and recurrence (Biber, 2009; Römer, 2009; Sinclair, 1991; Stefanowitsch & Gries, 2003), that is they appear together more often than would be "expected on the basis of chance" (Gries, 2008, p. 6) or they are very often repeated together (Biber et al., 1999; Paquot & Granger, 2012). Importantly, these studies also show that associations between lexical items and grammatical structures often serve specific communicative functions (Biber et al., 2004), which means that associations between lexis and grammar also need to be considered within the specific communicative context in which they occur. The problem with strictly separating lexis and grammar on scoring rubrics is therefore two-fold.

First, it ignores important information about linguistic competence encoded at the lexis-grammar interface, information which has been shown to contribute to holistic evaluations of learner productions. For example, Paquot (2019) found that the association between words in specific grammatical structures (e.g., arouse + curiosity in a direct object structure) was a significant predictor of the proficiency scores assigned to advanced learner texts in a corpus of research papers. Importantly, no significant differences were found regarding the lexical or syntactic complexity of the texts, suggesting that the most important information about proficiency was not lexical or grammatical but rather lexicogrammatical. Similarly, Kyle and Crossley (2017) found that the association strength of verbs and the constructions in which they appear (e.g., subject-**require**-direct object-direct object, 'sports require teamwork and the ability to cooperate') was likewise a predictor of holistic proficiency scores and that models based on such

lexicogrammatical phenomena were better predictors of proficiency than a model based solely on syntactic complexity. These results show that rating scales that do not account for lexicogrammatical phenomena may not be fully representative of the larger construct of proficiency, which constitutes a threat to the *explanation inference* in an argument-based approach to test validation (Chapelle et al., 2008; Kane, 2006).

The second problem with separating lexis from grammar on scoring rubrics is that it ignores the way that both lexis and grammar interact together with communicative function. That language varies according to context of use is certainly not news to language assessment specialists (see Bachman, 1990). What is less clear however, is the extent to which the functional differences between tasks interact with proficiency scores when it comes to lexicogrammatical phenomena. In a study comparing the most frequent multi-word units across various academic corpora, Biber, Conrad and Cortes (2004) for example, found a strong relationship between grammatical structure, discourse function and the types of texts where such units were found (e.g., classroom teaching vs. academic prose). However, as Paquot et al. (2021, p. 229) point out, "there is a need for more research into how lexicogrammatical competence interacts with foreign language proficiency and development across modes, tasks, registers and other linguistic, contextual and or situational characteristics." In particular, relatively little research has been done into the linguistic differences between oral and written production tasks on language exams (cf. Biber et al., 2016). The current study therefore represents an attempt to complement the discussion about the division between lexis and grammar in language testing by investigating the extent to which lexicogrammatical characteristics are predictive of proficiency scores in oral versus written exam tasks.

6.2 Background

One way to capture complexity at the lexis-grammar interface is through phraseology. Phraseology looks beyond single words to examine the way that words combine together. Psycholinguistic re-

search has shown that both learners and native speakers are sensitive to patterns of co-occurrence and recurrence but in different ways. Ellis, Simpson-Vlach and Maynard (2008), for example, found that for learners of English, the processing of three-, four- and five-word phrases is primarily a function of the frequency of the phrases in large reference corpora. For native speakers however, it is not frequency but rather association strength of the phrases that has the strongest effect on processing. Phrases containing words that occur more exclusively together (and rarely with other words) were the fastest to process. These receptive differences are also found in production data. Durrant and Schmitt (2009) compared written texts produced by native and non-native writers of English and found that compared to the native writers, the non-native writers used more highly frequent collocations and fewer infrequent but strongly associated collocations. Similarly, as learners increase in proficiency, they tend to decrease their use of highly frequent bigrams (adjacent word pairs) and increase their use of infrequent but highly associated bigrams, both when learners are compared cross-sectionally across proficiency levels (Granger & Bestgen, 2014) as well as when individual learners are followed longitudinally (Bestgen & Granger, 2018).

Paquot (2019) termed this aspect of linguistic competence *phraseological complexity*, which she defined as the diversity and sophistication of phraseological units in learner production (following Ortega, 2003). Phraseological diversity refers to the extent to which a learner uses a large number of different phraseological units. Phraseological sophistication on the other hand, refers to the extent to which a learner is able to select more specific or informative phraseological units (i.e., those which may be more appropriate to a specific topic or register). In a series of studies focusing on L2 English, French and Dutch, measures of phraseological complexity have been found to be predictive of L2 proficiency, even when lexical and syntactic complexity measures were not (Paquot, 2018, 2019; Rubin et al., 2021; Vandeweerd et al., 2021). While these studies speak to the relevance of phraseological complexity to L2 proficiency in general, they are limited by the fact that they have focused exclusively on written production.

Writing and speaking differ in a number of important ways including the presence (or absence) of an audience, the degree of online planning required and control over output (Ravid & Tolchinsky, 2002). The fact that speaking occurs online and in real-time also means that learners have less time to conceptualize, formulate and monitor linguistic output as compared to writing (Skehan, 1998). Even the most pressured writing conditions (e.g., timed language exams) still offer more opportunities for planning and revising compared to speech. As a result, L2 speech generally exhibits less lexical complexity as compared to writing (e.g., Granfeldt, 2007). The constraints of real-time production and the lack of online planning may also lead to the use of more frequent (and therefore more accessible) phraseological units, as suggested by the results of Biber and Gray's (2013) study of the collocations used in oral and written TOEFL exam tasks.

In addition to the overall differences outlined above, speech and writing often serve different communicative purposes. As a result, oral and written production tend to be associated with the use of different linguistic features (Biber & Conrad, 2019). For instance, the fact that speech is often interactive means that it tends to have a more personal or involved focus. Linguistically, this is associated with a greater use of pronouns, modal verbs, lexical verbs, adverbs and discourse markers. In contrast, writing tends to be more informational in purpose and is marked by the use of nouns, nominalizations, attributive adjectives and prepositional phrases. These differences have been observed across several languages and different text types (Biber, 2014) and even extend to the level of phraseology, with speech being associated with the use of verb-based phraseological units (lexical bundles) and writing being associated with noun-based phraseological units (Biber et al., 2004).

From a developmental perspective, linguistic features associated with speech are also hypothesized to be acquired early because they are necessary for almost all types of conversation. Linguistic features common to writing (e.g., complex phrasal embedding) on the other hand, are hypothesized to be acquired later, and only in the context of formal education (Biber et al., 2011; Halliday & Matthiessen, 2006). This suggests that although tasks may require different linguistic features,

there may be developmental patterns within tasks whereby more advanced language users incorporate more nominal complexity features, regardless of task type. For example, in a study comparing oral and written TOEFL exam tasks, Biber, Gray and Staples (2016) found that although written tasks were indeed associated with the use of phrase-level linguistic features and oral exams were associated with clause-level linguistic features, within each task, there was a positive relationship between the use of phrasal features and holistic proficiency scores given by raters.

With regards to French, the target language of the learners in our study, phraseological complexity has also been found to increase with proficiency both in oral production (Forsberg, 2010) as well as written production (Vandeweerd et al., 2021). Only one study thus far has directly compared phraseological complexity across modes however. Vandeweerd et al. (in revisions) (Chapter 5 in this thesis) compared the diversity and sophistication of direct objects (verb + noun) and adjectival modifiers (noun + adjective) across oral and written tasks collected from university-level learners of French over a period of 21 months. While the design of the corpus made it difficult to isolate the precise effects of task characteristics (e.g., modality, interactivity) from register, comparison of the learner group and an L1 group seemed to indicate that the extra planning time afforded by the written task had a positive effect on the phraseological complexity in learner productions but that this effect was somewhat mediated by the functional requirements of the task, as seen by the fact that the complexity of the noun-based units (adjectival modifiers) and verb-based units (direct objects) differed across tasks. However, the applicability of these findings to the language testing context is somewhat limited given that the study involved a relatively small and quite advanced sample learners (n = 29) and the fact that the learners were not directly impacted by the results of their linguistic performance.

The current study therefore aims to determine the extent to which measures of phraseological complexity are predictive of oral versus written French proficiency scores in the context of a high-stakes language assessment exam. In line with previous research, we expect that measures of phraseological complexity will be predictive of

proficiency even when controlling for the lexical complexity of a text (Vandeweerd et al., 2021). This would suggest that raters are sensitive to complexity not just at the level of lexis but also at the lexis-grammar interface (i.e., that phraseological measures are not simply a subset of lexical complexity measures) as argued by Paquot, Gries, et al. (2021). We also aim to determine the extent to which various measures of phraseological complexity (e.g., those targeting diversity versus sophistication and noun-based versus verb-based units) differ in their ability to predict proficiency scores in the two modes. We expect to find that verb-based measures will be more predictive of proficiency in the oral task and that noun-based measures will be more predictive of proficiency in the written task due to the register-induced differences between the two tasks (in line with Biber et al., 2016) but that the constraints of online production will have a mediating effect on the complexity of speech (Skehan, 1998).

6.3 Methodology

6.3.1 Data

This study uses a sample of 545 oral and written productions from the *Test d'évaluation de français* (Chambre de commerce et d'industrie de Paris, 2010), a French proficiency exam recognized by the French Ministry of Education for the admittance of non-native speakers to higher level education in France. Because the exam is also recognized by the federal government of Canada for immigration purposes, many of the candidates who take the exam are themselves native speakers of French. We have included both L1 and L2 data (using L1 background as a control variable in the statistical analysis). The focus of the current study is on two of the production tasks:

- Written Production Task B asks candidates to write a 'letter to an editor' (min. 200 words) in which they respond giving their opinion on a controversial topic (e.g., abolishing television advertisements). Candidates have one hour to complete this task as well as another written production task. Both tasks are completed via computer without access to reference materials.

- Oral Production Task B asks candidates to convince a close friend to participate in an activity. They are provided with a prompt (e.g., an advertisement promoting a vacation to Quebec) and must 'telephone' the examiner who acts as their friend. Candidates have ten minutes to complete the task.

Table 6.1: Sample of texts used from TEF corpus

	oral, N = 545	written, N = 545
L1		
Arabic	226 (41%)	226 (41%)
French	173 (32%)	173 (32%)
Haitian Creole	11 (2.0%)	11 (2.0%)
Kinyarwanda	12 (2.2%)	12 (2.2%)
other	123 (23%)	123 (23%)
TOKENS	966 (145, 2,006)	271 (104, 700)
SCORE	551 (263, 699)	475 (165, 699)

[1] n (%); Median (Range)

These two tasks were chosen because of their similarity in terms of communicative function. Specifically, both tasks require candidates to express their personal attitudes and their certainty about propositions in order to persuade the addressee (i.e., the newspaper editor in the case of the written task and the 'friend' in the case of the oral task).

The evaluation of the tasks was done by two independent raters who judged the production on scales of 0-6 according to how well the candidate fulfilled the task (e.g., presentation of ideas, quality of argumentation) as well as linguistic criteria including grammatical precision, vocabulary control, elocution (in the case of the oral task) and orthography (in the case of the written task). The final score for each subcriterion was the result of the mean score of both raters. A formula was then used to transform the individual scores into a normed score out of 699 for the entire section of the exam, with 100 point increments indexed to the six levels of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001; Demeuse et al.,

2010).

The sample of exams used in the study (Table 6.1) was made according to several criteria. First, we made an initial selection of exams where the written section was at least 100 words in length given that complexity measures have been shown to be unreliable for texts shorter than that. Second, given that topic can have a significant effect on phraseological complexity (Paquot, Naets, et al., 2021), we wanted to ensure that differences observed between oral and written sections of the exam were not simply due to topic differences. We therefore manually categorized the prompts into overarching themes (e.g., environment, travel, language, etc.) and then made an initial selection of exams with a match between the theme of the oral and written sections. The audio from the oral exams was then transcribed (see supplementary materials for transcription guidelines) and the written exams were manually spell-corrected in order to ensure comparability with the oral transcriptions (which by definition did not have any orthographic errors). Letter headings (e.g., addresses, signature lines) were also removed from the written productions. In house R scripts (R Core Team, 2021) were then used to clean the texts (e.g., remove stylized apostrophes, standardize spelling) and part-of-speech (POS)-tag them using Treetagger (Schmid, 1994).

Following pre-processing, we used topic-modeling as an additional method of controlling for topic differences. Topic-modelling is a bottom-up method whereby texts are grouped together into computer-defined topics according to their common vocabulary (see Murakami et al., 2017). For each text in the corpus, the model outputs a probability distribution for each topic. This allows the differences between texts to be quantified in terms of their overlapping probability distributions using measures such as Jensen-Shannon Divergence (JSD). The smaller the JSD between two texts, the more similar they are in terms of shared vocabulary. We used the JSD scores between the oral and written sections of an exam to control for topic differences in the statistical analysis. The full topic-modelling procedure is described in the supplementary materials.

6.3.2 Complexity measures

Previous studies of phraseological complexity have primarily focused on co-occurrence phenomena (e.g., adjective + noun, verb + noun, adverb + verb collocations). However, equally important to L2 phraseology are recurrence phenomena, such as *lexical bundles*, defined as "simple sequences of words that commonly go together in natural discourse" (Biber et al., 1999: 990). Looking at lexical bundles means that a broader set of phraseological phenomena can be captured. This is especially relevant for French because as noted by Vinay and Darbelnet (1995: 126), French has a tendency to modify nouns and verbs with prepositional phrases. For example, the French equivalent of *comment ironically* would be *s'exprimer en termes ironiques* (lit. 'express oneself in ironic terms'). Looking only at adverb + verb collocations would ignore this type of verbal modification (as argued in Vandeweerd et al., 2021).

We therefore decided to follow Biber, Conrad and Cortes (2004) in using a bottom-up method to identify phraseological units rather than focusing on specific syntactic structures. We did however make some modifications to the way that lexical bundles are traditionally extracted. Whereas lexical bundles are typically composed of inflected surface forms, in our approach we used POS-tagged and lemmatized lexical bundles. Using POS-tagged lexical bundles allowed us to filter the units we extracted in order to account for possible cross-modal register differences (Biber et al., 2004). We focused on two types of lexical bundles in particular: those starting with verbs (verb-bundles) to tap into phraseological complexity at the clause-level and those starting with nouns (noun-bundles) to tap into phraseological complexity at the phrase-level. Admittedly, this is a rather crude operationalization but given that this is the first study to investigate phraseological complexity in this way, we felt that it represented a fair balance between a strictly bottom-up approach and imposing too much of our own intuition on the units extracted. The decision to use the lemmas instead of the surface forms was due to the highly inflected nature of French and follows the suggestion of Treffers-Daller (2013) to use lemmas in the calculation of complexity measures in French. In addition, this protects against a

possible confound of accuracy in the calculation of our phraseological complexity measures because all inflected forms are treated equally. This is especially important when comparing oral and written French because many grammatical inflections are only encoded orthographically but not phonologically (see Blanche-Benveniste & Adam, 1999). On a practical level, collapsing inflected forms together also makes it possible to calculate association measures in cases where an exact inflected form may not be attested in a reference corpus but another inflected form, representing a different tense or mode *is* attested, thereby increasing the coverage of our phraseological complexity measures. Another difference from traditional lexical bundle approaches is that we included bundles that crossed punctuation (but not sentence) boundaries (cf. Biber & Barbieri, 2007). This was done because the oral transcriptions contain no punctuation (other than the segmentation into C-units) so we wanted to ensure comparability across modes.

All lexical bundles were extracted using R scripts. Following extraction, the list of units was filtered to remove any verb-bundles starting with the auxiliary verbs *être* and *avoir* (so that the focus was on lexical verbs) and any bundles containing lemmas unknown to Treetagger. We then collapsed all proper nouns and determiners within the bundles. Without collapsing proper nouns, many of the units are not found in the reference corpus (simply because the specific names are not attested). Collapsing determiners was also done out of practical necessity because otherwise indexing all variations of units with different determiners is no longer feasible. The frequency of each bundle was then checked in a large reference corpus and those that appeared fewer than ten times were excluded from analysis. The reference corpus used was the FRCOW16 corpus (Schäfer, 2015; Schäfer & Bildhauer, 2012): a ten billion word web-scraped corpus of French. An obvious disadvantage of this corpus is that it only contains written texts. As such, it may not be exactly representative of the tasks completed by TEF candidates. That being said, the size of the corpus means that a majority of the lexical bundles extracted from the TEF exams are attested, in comparison to smaller oral-specific corpora in which many of the extracted lexical bundles were simply not found (for a similar argument

Table 6.2: Most frequent bundles extracted from TEF corpus

Bundle	Example	Type	PMI	Tokens			
				TEF Written	TEF Oral	TEF Total	FRCOW
penser_VER que_KON ce_PRO suivre|être|sommer_VER	*je pense que c'est positive*	verb-bundle	6.27	5	368	373	91426
monsieur_NOM DET rédacteur_NOM en_PRP	*monsieur, le rédacteur en chef*	noun-bundle	5.62	135	0	135	137
lecteur_NOM de_PRP DET journal|journau_NOM	*lecteur de votre journal*	noun-bundle	8.48	114	0	114	187
faire_VER part_NOM de_PRP DET	*faire part de mon point de vue*	verb-bundle	5.87	100	0	100	59530
dire_VER que_KON ce_PRO suivre|être|sommer_VER	*Je tiens à vous dire que c'est absurde*	verb-bundle	5.53	3	84	87	138227

Note:
Vertical bars indicate lemmas that could not be disambiguated by Treetagger and were therefore collapsed in the analysis.

see Paquot, Naets, et al., 2021). As a precaution, we checked whether there was a significant difference in coverage of lexical bundles from the oral and written tasks and this was not found to be the case. Table 6.2 shows the most frequent bundles extracted from the TEF corpus.

Following Paquot (2019), we defined phraseological complexity in terms of the diversity and sophistication of the bundles. Whereas previous phraseological complexity research has operationalized diversity using transformations of type-token ratios, the length of the phraseological units in the current study mean that very few of them are repeated verbatim more than once in a given text. As a result, type-token based measures are not very informative. Instead, we decided to operationalize diversity as the ratio of bundle types (e.g., *penser_VER que_KON ce_PRO être_VER*, 'think that it is') to the number of different POS structures ('VER KON PRO VER'). The logic behind this measure is that a learner who has a relatively small phraseological repertoire will consistently use the same words in the same grammatical structures whereas a learner with a larger phraseological repertoire will show more variation in the words they use to fill grammatical structures and so the measure is indicative of a learner with a broader phraseological repertoire. Phraseological sophistication was operationalized as the pointwise mutual information (PMI, Church & Hanks, 1990) between the first word in the

bundle (e.g., *penser_VER*) and the rest of the sequence (e.g., *que_KON ce_PRO être_VER*). This represents the extent to which the first word (the noun or the verb) *selects* the rest of the sequence (Dunn, 2018). PMI highlights associations that are highly exclusive. Because of this exclusivity, units with a higher PMI tend to serve specific discursive functions (N. C. Ellis et al., 2008). For example, *salutation_NOM DET plus_ADV distingué_ADJ* ('most distinguished salutation'; PMI = 18.5), is used as a greeting in formal letters. In contrast, bundles with a lower PMI tend to be more general and part of everyday use by virtue of the fact that the noun or verb is often found in other structures (e.g., *trouver_VER que_KON ce_PRO suivre|être|sommer_VER*; 'find that it's'; PMI = 4.5).

Because of the text length differences between the oral and written productions (see Table 6.1), we decided not to use measures based on the entire texts but to use a random sampling method instead. Both measures of phraseological complexity were calculated as the average over ten random samples of five bundles.[1] Visual inspection of the data revealed that the diversity measures exhibited positive-skew, which was addressed by applying a Box-Cox transformation to both measures. In addition to the phraseological complexity measures, we calculated two measures of lexical complexity. Lexical diversity was operationalized using the Measure of Textual Lexical Diversity (MTLD, McCarthy & Jarvis, 2010), and lexical sophistication was operationalized as the number of infrequent words (words on the 3-, 4- and 5-thousand frequency bands of the FRCOW16 corpus), averaged across 100-word moving windows.

6.3.3 Analysis

We built a mixed effects regression model using the *nlme* package (Pinheiro et al., 2020) in R in order to predict the proficiency scores of the exams as a function of the lexical and phraseological complexity measures. Exams with fewer than five noun-bundle or verb-bundle types were excluded as well as exams where topic could not be con-

[1] The sample size of five was chosen because only 13 exams had fewer than five noun-bundles or verb-bundles in either production. These exams were excluded from analysis.

clusively assigned by the model (n = 51). The following variables were included as controls in the model: L1 of the candidate, JSD (the topic similarity between the oral and written sections of exams) and the total number of words (tokens) to control for text length. Because each candidate contributed two productions to the dataset, the model included random intercepts per candidate ID. We initially began with a fully-specified model with all levels of L1 and topic (the 7 model-generated topics) and where each of the lexical and phraseological complexity measures were allowed to interact with both mode (written/oral) and topic. We then used backwards step-wise model selection in order to determine if all levels L1 and topic were necessary to the model and whether the interactions between mode and topic were necessary for each complexity measure (i.e., whether they significantly improved model fit). After simplifying interactions, non-significant predictors were retained in the model if they either acted as a control variable (e.g., JSD) or if they were necessary to answer the research questions (in which case a non significant result would indicate that the null hypothesis that the given measure is not predictive of proficiency could not be rejected). All numeric predictors were converted to z-scores prior to modelling. Initial models revealed problems with homoscedasticity which were resolved by weighting variance according to mode. In order to evaluate the generalizability of the model, the data were split into a training set (67%) for model building and a test set (33%). Following the model selection process, only five levels of L1 (French, Arabic, Haitian Creole, Kinyarwanda and other) and two levels of topic (travel and other) were retained in the model. We also tested whether 2nd and 3rd degree polynomial transformations of the complexity measures would improve model fit in order to determine if any of the measures had non-linear relationships with proficiency. This was only found to be the case for MTLD, for which a 2nd degree polynomial significantly improved model fit.

6.4 Results

The summary of the multiple regression model is provided in Table 6.3. This model was able to explain 47.21% of variance in proficiency scores

Table 6.3: TEF model summary

Fixed Effects	b	95% CI	SE	df	t	p
(Intercept)	559.658	[545.39,573.927]	7.28	359	76.879	<0.001
L1ara	-42.35	[-56.025,-28.674]	6.977	359	-6.069	<0.001
L1hat	-6.573	[-49.23,36.085]	21.764	359	-0.302	0.763
L1kin	-16.252	[-56.868,24.363]	20.722	359	-0.784	0.433
L1other	-59.504	[-75.841,-43.167]	8.335	359	-7.139	<0.001
JSD	-4.352	[-10.177,1.473]	2.972	359	-1.464	0.144
TOKENS	55.705	[47.63,63.78]	4.12	353	13.521	<0.001
poly(MTLD, 2)1	690.68	[475.113,906.247]	109.983	353	6.28	<0.001
poly(MTLD, 2)2	-157.539	[-296.07,-19.008]	70.679	353	-2.229	0.026
LEX.SOPHISTICATION	0.972	[-5.714,7.658]	3.411	353	0.285	0.776
TOPIC.travel	-2.456	[-15.853,10.942]	6.835	353	-0.359	0.720
PHRAS.SOPHISTICATION.N	5.291	[-0.633,11.215]	3.023	353	1.751	0.081
PHRAS.SOPHISTICATION.V	7.768	[1.238,14.297]	3.331	353	2.332	0.020
PHRAS.DIVERSITY.N	-4.694	[-11.347,1.958]	3.394	353	-1.383	0.168
MODEwritten	-28.372	[-49.888,-6.855]	10.978	353	-2.584	0.010
PHRAS.DIVERSITY.V	0.036	[-5.474,5.547]	2.811	353	0.013	0.990
LEX.SOPHISTICATION:TOPIC.travel	20.664	[3.259,38.07]	8.88	353	2.327	0.021
PHRAS.DIVERSITY.N:MODEwritten	14.059	[3.898,24.22]	5.184	353	2.712	0.007
Random Effects	**Variance**	**sd**				
(Intercept)	1603.769	40.047				
Residual	6253.346	79.078				

Note:
JSD = Jensen Shannon Divergence
MTLD = Measure of Textual Lexical Diversity
Call: SCORE = L1 + JSD + TOKENS + poly(MTLD, 2) + FRCOWOFF2K * TOPIC + DIV.PMI.N + DIV.PMI.V + DIVERSITY.N.bcn * MODE + DIVERSITY.V.bcn
Marginal R^2:0.337
Conditional R^2:0.472

overall (conditional R^2) and 33.66% of the variance was explained by the fixed effects in the model (marginal R^2). On unseen data, the model was able to explain 36.81% of the variance in proficiency scores. Both the residual mean squared error (73.23) and mean absolute error (58.9) indicated that when the model was off, it was usually off by less than one hundred points (i.e., less than one CEFR level). Model diagnostics did reveal some non-linearity of residuals but this was largely due to data scarcity at the lower end of the proficiency scale. For this reason, the accuracy of the model for proficiency scores lower than 400 (CEFR B2) should be interpreted with caution.

In the two sections that follow, we present the results of the model concerning the lexical and phraseological complexity measures. As a reminder, these results show the effects of each measure, while controlling for L1 background, text length and topic similarity between the oral and written components of exams.

6.4.1 Lexical complexity measures

As shown in Figure 6.1, there was a significant curvilinear effect of lexical diversity (MTLD) on exam scores whereby an increase in one standard deviation of MTLD was found to be associated with approximately 27.26 (95% CI = 19.5,35.01) extra points on the exam but this plateaued at z > 1, after which an increase in lexical diversity was no longer associated with an increase in test scores. Lexical sophistication was also found to be positively correlated with test scores but only for those texts classified as belonging to the topic of travel (as shown by the significant interaction between lexical sophistication and topic). For texts on the topic of travel, an increase in one standard deviation was associated with an increase of 21.64 (95% CI = 1.09,42.18) points on the exam. The lexical sophistication of texts not classified as belonging to the topic of travel was found to be associated with an increase in 0.97 (95% CI = -7.23,9.18) points (n.s.).

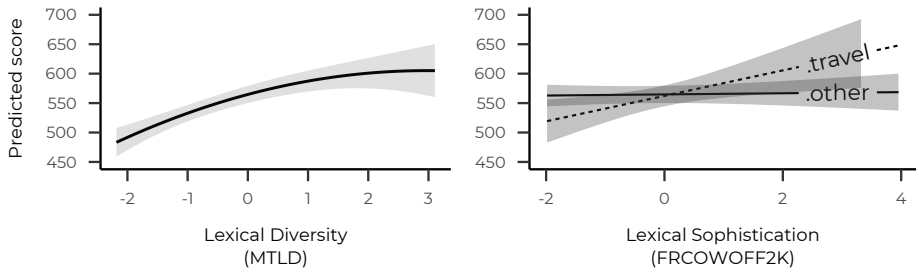


Figure 6.1: Effects plots for lexical complexity measures

6.4.2 Phraseological complexity measures

The effects plots for all phraseological complexity measures are shown in Figure 6.2. None of the phraseological complexity measures were found to interact significantly with topic and only one of the phraseological complexity measures was found to interact significantly with mode: the diversity of noun-bundles. In the written task, a one standard deviation increase in the diversity of noun-bundles was associated with 9.36 (95% CI = -0.16, 18.89) extra points on the exam but in the oral mode, this was associated with 4.69 (95% CI = -12.86, 3.47)

fewer points on the exam. The diversity of verb-bundles was found to be slightly positively correlated with proficiency scores but this was not significant (EMM trend = 0.04, 95% CI = -5.49, 5.57). In terms of sophistication, positive correlations were found with proficiency for both noun-bundles (EMM trend = 5.29, 95% CI = -0.65, 11.24) and verb-bundles (EMM trend = 7.77, 95% CI = 1.22, 14.32) in both modes but only the correlation with the sophistication of verb-bundles was found to be significant at an alpha level of 5%.

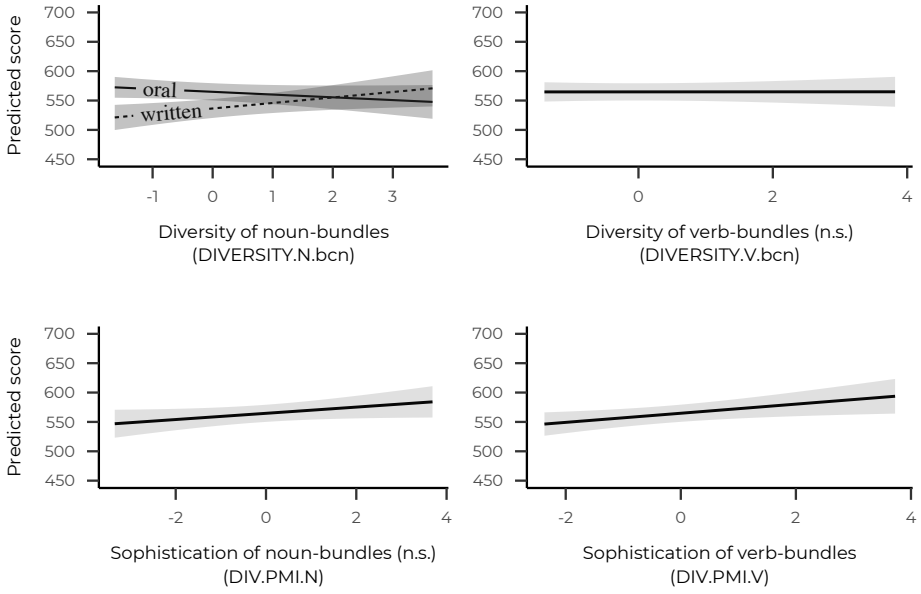


Figure 6.2: Effects plots for phraseological complexity measures

6.5 Discussion

This study aimed to determine: (1) the extent to which measures of phraseological complexity are predictive of oral versus written proficiency scores and (2) the extent to which various measures of phraseological complexity differ in their ability to predict proficiency scores in the two modes.

A mixed effects regression model revealed that two of the four phraseological complexity measures were predictive of proficiency scores,

Table 6.4: Comparision with alternative models

Model Type	AIC	Marginal R^2	RMSE
Lexical and phraseological	8310.253	0.368	73.227
Only phraseological	8355.128	0.344	74.626
Only lexical	8318.003	0.350	74.259

Note:
AIC = Akaike information criterion
RMSE = Root mean squared error

namely the diversity of noun-bundles (in the case of the written productions) and the sophistication of verb-bundles (in both the written and oral productions). Importantly, these correlations held even when controlling for the lexical diversity and sophistication of texts. As a comparison, we also built two alternative models, the first containing only lexical measures and the second containing only phraseological measures. As shown in Table 6.4, both alternative models had poorer model fit (AIC) than the model with both lexical and phraseological measures ($p < 0.001$), explained less variance in test scores (Marginal R^2) and had slightly more error in predictions made on unseen data (RMSE). This indicated that both types of measures were important predictors of proficiency scores and that lexical complexity measures did not increase their explanatory power once the phraseological measures were removed from the model.

That being said, phraseological measures were found to exhibit two characteristics that made them especially useful as indicators of proficiency. First, unlike with lexical diversity, there was no plateau in the relationship between phraseological complexity and proficiency scores. Increases in phraseological complexity continued to contribute to proficiency scores even at the highest levels of proficiency. Second, none of the phraseological measures were found to interact significantly with topic (cf. Paquot, Naets, et al., 2021 who found a significant effect of topic on phraseological sophistication). This was not the case for lexical sophistication, for which the correlation was only significant for texts on the topic of travel. That being said, it is important to highlight that the contribution of phraseological

complexity to exam scores amounted to fewer than 10 extra points on the exam for every one standard deviation increase in phraseological complexity (holding all other variables constant). While this is a rather modest contribution, it is still much larger than that of lexical sophistication, which for texts not on the topic of travel, only amounted to an increase of about 1 extra point (n.s.) for every one SD increase in lexical sophistication. This, despite the fact that examiners are explicitly asked to evaluate vocabulary use in the scoring rubric for the TEF exam.

The results of this study also showed that measures of diversity and sophistication differed in their ability to predict proficiency scores. In both written and oral productions, the *diversity* of noun-bundles was more predictive of proficiency than the sophistication of noun-bundles, suggesting that all else being equal, what mattered most to raters was not how sophisticated noun-bundles were but rather the diversity in the words used to fill noun-bundle structures. In the case of written productions, using a diverse set of noun-bundles was associated with higher proficiency scores. This is likely because noun-bundles tap into the phrasal complexity features typical of academic, written discourse (Biber & Conrad, 2019; Halliday & Matthiessen, 2006). A candidate who uses a variety of noun-bundles in written production therefore demonstrates both knowledge of the functional requirements of the register (e.g., the use of nominalizations and post-modification of nouns) and shows evidence of a broad lexicogrammatical repertoire.

In oral productions on the other hand, greater noun-bundle diversity was associated with lower proficiency scores. This may be because phrase-level complexity features are less important in speech, which tends to have a more personal or involved focus. As such, noun-bundle diversity in speech may demonstrate poor awareness of register conventions. Another possibility is that the "diversity" of oral productions is actually due to the repetition of a small set of semi-fixed units, in line with Römer's (2017) finding that spoken corpora are made up of highly frequent but variable expressions (e.g., I think/mean/thought/know/suppose it) and Leech's claim (2000, p. 697) that speech relies on a "restricted and repetitive lexicogrammat-

ical repertoire" due to the constraints of real-time processing. For example, in one low-scoring oral production (CEFR B1) a candidate used two noun-bundles with the same final three words: *temps avec tes enfants* ('time with your children'; PMI = 4.5) and *sport avec tes enfants* ('sport with your children'; PMI = 3.05) within the same utterance. Although the two bundles are technically unique and contribute towards overall noun-bundle diversity, a rater may have been put off by the repetition of the last three words in close proximity. To check if this might the case more broadly, we calculated the number of noun-bundles in each text that differed by only one word from another noun-bundle in the same text. As suspected, noun-bundles were repeated with only a one-word difference 0.55 (SD = 0.94) times on average in oral productions, compared to 0.18 (SD = 0.45) times in written productions. Of course, this is only a very crude estimate (and there is a large amount of variation) but it hints at the potential usefulness of looking at patterns of lexicogrammatical "fixedness" in French, that is, the use of fixed, invariable expressions compared to "slot-and-frame" expressions with variable internal components (along the lines of Biber's 2009 study of such patterns English). At present, we can only speculate on potential reasons for the cross-modal differences observed in the data and there are likely other confounding factors (e.g., communicative success, accuracy, fluency) playing a role. That being said, what *is* clear is that when it comes to lexicogrammatical diversity, simplistic descriptors such as *the more…the better* are not necessarily accurate for oral production.

When it comes to verb-bundles, the results showed that it was *sophistication* rather than diversity that was most predictive of proficiency scores. Although we had expected that verb-based measures would be more predictive of proficiency in the oral productions than in the written productions, we found no significant interaction effect of mode. The sophistication of verb-bundles was positively associated with proficiency to the same extent in both written and oral productions. In retrospect, this finding is actually quite logical because both tasks require argumentation: an opinion about a controversial topic in the case of the written task and reasons why a friend should participate in an activity in the oral task. As shown by Michel et al. (2019),

argumentative tasks tend to be characterized by high levels of subordination compared to other types of tasks. Moreover, lexical bundles that are used to express attitudes or assessments of certainty (what the Biber et al. (2004) refer to as 'stance bundles') tend to be composed of dependent clauses and verb phrases (roughly equivalent to the verb-bundles in the current study). This means that the argumentative aspect of both tasks (persuading the addressee) requires the use clause-level syntactic features and so the major factor differentiating candidates is therefore the *sophistication* of the verb-bundles. Candidates who use more sophisticated verb-bundles show evidence of having a deeper lexicogrammatical repertoire (all other things being equal). This finding is consistent with the longitudinal results of Vandeweerd et al. (2021) (Chapter 5 in this thesis), who found that the sophistication of verb + noun collocations increased significantly over a 21-month period in both written and oral tasks produced by university-level learners of French but that no significant change was observed for adjective + noun collocations.

Together with the findings of Vandeweerd et al. (2021), our results suggest that because verb-bundles tap into communicative functions that are common to both oral and written registers, they may be more useful as broad measures of lexicogrammatical competence when comparing across modes. Moreover, the fact that these measures were predictive of proficiency scores even though raters of the TEF were not explicitly asked to evaluate this aspect, suggests that lexicogrammatical complexity at the verb-level is particularly salient. This may be because verb-bundles tap into communicative events that are common across all exams, facilitating comparisons between candidates. For example, in the case of the written task, candidates must introduce the topic of their letter and often rely on specific expressions to do so. The difference between *je me permets de réagir à* ('I *permit* myself *to react to*,' PMI = 6.19), and *je vous écris pour donner mon avis* ('I *write* you *to give my* opinion,' PMI = 2.60) provides a clear opportunity to compare candidates. Although both expressions are grammatically well-formed, the first one is more indicative of the formal register required for a letter to a newspaper editor. Likewise, in the oral task, candidates must find a way to

introduce the 'opportunity' they have for their friend. A rater may consider the use of *tombé sur l'annonce* ('fell on the advertisement,' PMI = 9.09) to be more indicative of higher proficiency because it shows evidence that the candidate is aware of the informal nature of the oral task (conversation between friends) in comparison to the more neutral expression, *vu la publicité pour* ('saw the advertisement for,' PMI = 3.79). When providing training to raters, it may therefore be useful to highlight communicative events such as 'introducing the purpose of the letter' as potential sources of information about a candidate's lexicogrammatical competence, especially given that raters already seem to be somewhat naturally attuned to them.

Taken together, the results of this study suggest that lexicogrammatical complexity represents a separate construct from that of lexical complexity and provide more evidence in favour of including lexicogrammatical criteria within language testing rubrics (Paquot, Gries, et al., 2021). In Europe, proficiency tests such as the TEF are indexed to the Common European Framework of Reference (CEFR, Council of Europe, 2001). As Paquot (2018) points out, the CEFR promotes a rather traditional understanding of phraseology and does not do justice to the ways that both lexis and grammar interact with communicative function, especially at the higher proficiency levels. As an example, in both the original descriptors and the revised version (Council of Europe, 2018), the use of "prefabricated expressions" is listed as an indication of lower proficiency. This contrasts somewhat with the results here, which found that certain lexical bundles are actually quite sophisticated by virtue of their exclusivity (PMI) and the fact that they demonstrate knowledge of the characteristics and linguistic requirements for certain registers (e.g., the use of *je me permets de réagir* in formal written letters). To be fair, the most recent version of the CEFR *does* acknowledge the importance of co-occurrence and recurrence phenomena but this is only within the description of *vocabulary control* in general, not within the descriptor scales: "as competence increases, such ability is driven increasingly by association in the form of collocations and lexical chunks, with one expression triggering another." (Council of Europe, 2018: 134). However, as shown by the results of this study, reducing the relationship

between lexicogrammatical competence and proficiency to a simple linear association, as has previously been done in phraseological complexity studies, does not do justice to the complex ways that lexis and grammar both interact with register, modality and proficiency.

6.6 Conclusion

In the introduction to this article, we argued (in line with Römer, 2017) that conceptualizing lexis and grammar as solely separate phenomena may lead to invalid inferences about a candidate's linguistic competence given evidence from target language use. We also discussed the ways in which phraseological phenomena are likely to differ between written and spoken exam tasks. Specifically, we hypothesized that the association between phraseological complexity and proficiency would be mediated by constraints of online processing (Skehan, 1998) and register (Biber & Conrad, 2019). In order to investigate this, we presented the results of a study in which we compared the complexity (diversity and sophistication) of two types of phraseological units (noun-bundles and verb-bundles) between oral and written productions from the TEF exam. The results of this study showed that phraseological complexity measures were significant predictors of the proficiency scores assigned to texts by raters, even when controlling for lexical complexity. As it turned out, these measures were even better predictors of proficiency than measures of lexical diversity or sophistication. The results also spoke to the interplay between phraseological complexity, modality and register when it comes to evaluating proficiency.

That being said, it is important to highlight the limitations of the current study. For one, in sampling the TEF exams, we made a decision to prioritize cross-modal topic similarity but this came at the expense of a less diverse sample of proficiency levels. This meant that the data that was used to build the model was skewed towards higher proficiency levels so the applicability of these findings to lower-level learners (below CEFR B2) will need to be further investigated in future studies. In addition, our rather broad focus on all noun- and verb-bundles in this study may be criticized for lack of linguistic precision. While it is cer-

tainly the case that not all units that were extracted were necessarily equally as informative about proficiency, this study does represent a good first step in broadening the scope of phraseological complexity measures beyond co-occurrence phenomena to include recurrence phenomena such as lexical bundles. In future studies, it may be fruitful to explore the differences between different noun or verb-bundle structures and to evaluate the usefulness of other aspects of recurrence (e.g., fixedness as in Biber, 2009). Finally, it is important to acknowledge that although lexicogrammatical complexity was shown to be relevant to the evaluations of proficiency in this study, there are of course many other pragmatic and linguistic factors that raters take into account when scoring an exam. Future studies would do well to explore the extent to which lexicogrammatical complexity works in combination with other aspects of proficiency (e.g., accuracy, fluency) as well as how both factors contribute to the ability to accomplish the exam tasks.

While we realize that a complete description of the complicated relationship between phraseological complexity, register and modality is probably beyond the scope of most scoring rubrics, the results of this study show that if descriptors are too simplistic, they risk painting an inaccurate picture of linguistic competence, which constitutes a threat to inferences about a candidate's proficiency made on the basis of exam results. At best, ignoring lexicogrammatical competence may under or over-estimate a candidate's proficiency. At worst, scoring rubrics may suggest associations between linguistic features that are actually negatively correlated with proficiency. Considering the results of this study, it seems clear that scoring rubrics should aim to include a separate criterion for lexicogrammatical competence, in addition to lexis and grammar because as argued by Paquot et al. (2021: 229), lexicogrammtical knowledge is not simply the "sum of lexical and grammatical knowledge." That being said, we agree with Römer (2017) that as corpus linguists, we do not have the relevant expertise to provide advice about the precise way that these findings should be applied. We do hope however that the results of the current study *can* help to inform a more empirically-grounded approach to the lexis-grammar interface in language testing.

Chapter 7

Discussion and Conclusion

Overview

This chapter returns to the three main research questions of this thesis and attempts to formulate answers based on the empirical studies presented in the preceding chapters. Before doing so, Section 7.1 briefly outlines some important similarities and differences between the studies that are important to keep in mind when comparing the results. Section 7.2 then addresses the first research question, namely the extent to which measures of phraseological complexity are predictive of proficiency and development in L2 French. Section 7.3 concerns the second research question, namely the comparison of phraseological complexity measures with 'traditional' measures of L2 complexity within lexical, syntactic and morphological domains. Section 7.4 attempts to answer the third research question which focuses on the differences in phraseological complexity in writing versus speaking. Section 7.5 then highlights some of the limitations of this thesis and points to some possible areas that could be explored in future research. Finally, Section 7.6 offers some concluding remarks.

7.1 Comparability of results across studies

Before turning to the research questions, it is important to mention that the three studies presented in this thesis were conducted using separate learner corpora and therefore contain data collected from different populations of language users. As such, care needs to be taken when attempting to draw parallels between them. With that in mind, there are some relevant similarities and differences that bear mention. In terms of learner populations, while both the LCF (Chapter 4) and the LANGSNAP corpus (Chapter 5) contain data from university-level learners of French, the learners whose data make up the LCF corpus are Dutch learners of French in Belgium, which in most cases constitutes a second language context, given the status of French as an official language in Belgium.[1] The participants whose data was collected for the LANGSNAP corpus, were for the most part L1 English speakers in the United Kingdom, most of whom had only experienced French in a foreign language context prior to their sojourn abroad in France. In comparison to both the LCF and LANGSNAP corpora, which both contain a rather homogeneous sample of learners, the TEF exam corpus used in Chapter 6 contains data from candidates all over the world. It is therefore made up of a wide variety of different L1 backgrounds and amounts of exposure to French. This does include a subset of data from university-level learners (those taking the exam to pursue higher education in France) but also includes a wide range of other types of learners as well as native speakers (those taking the exam to emigrate to Canada).

In terms of proficiency, as shown in Table 7.1, the texts in both the LCF and TEF corpus are skewed towards the higher CEFR levels but the TEF corpus also contains texts at a lower level than the LCF. No text-based proficiency scores are available for the LANGSNAP corpus but given that the learners had on, on average, 10.56 (SD = 3.01) years of previous education in French, it is reasonable to assume that they were also at a relatively advanced level, notwithstanding the methodolog-

[1] This is not meant to imply that all learners in this corpus had regular exposure to French in their daily lives. The amount of contact, of course, varies from any one learner to the next, especially as the universities were located in Flanders, a majority Dutch-speaking region. In fact, some scholars would argue that everyday exposure to French in Flanders is so limited as to be a 'foreign language context' (see Bulté et al., 2008).

Table 7.1: Proficiency (CEFR) distributions in LCF and TEF data

Mode	A1	A2	B1	B2	C1	C2
LCF						
written	-	-	-	26	106	37
TEF						
oral	-	1	26	115	275	128
written	1	13	109	204	168	50

ical difficulties of using years of education as a proxy for proficiency (Thomas, 1994). That being said, there was some variation in initial EIT scores (Mean = 62.9/120, SD = 17.97), which shows that the participants were not necessarily all at the same level prior to departure (see also Mitchell et al., 2017: 79). Following their sojourn abroad, this gap had reduced somewhat but there was still some degree of difference between participants (Mean = 85.62/120, SD = 15.18). Although it was not feasible within the context of this project, it would be useful for future research if the LANGSNAP texts were also externally assessed for proficiency (e.g., on the basis of CEFR levels) as this would facilitate comparisons across studies and would also serve as useful baseline of comparison with text-internal indices (e.g., complexity measures).

On the face of it, the three studies also differ with respect to the type of task that learners were asked to perform. The written task in both the LCF and the LANGSNAP corpus is an argumentative essay whereas the task in the TEF corpus is a 'letter to the editor.' Of course, essays and letters differ with respect to stylistic and structural conventions and are written with different addressees in mind. Yet, from a functional point of view, there are relevant similarities. In both tasks, learners need to communicate their point of view on a controversial topic, which requires, for example, the use of dependent clauses (e.g., *À mon avis, c'est vrai que*..., 'In my opinion, it's true that...'). Both the argumentative essays and the letters to the editor also have highly specific topics (e.g., nuclear energy, taxes on fast food, reducing pollution) which require the use of specialized vocabulary. There is less overlap between the oral tasks however. In the case of the LANGSNAP corpus, both the narrative and the interview focus on a progression of events

(see Section 5.4.1), whereas the telephone call task in the TEF corpus is somewhat more argumentative in that it requires candidates to convince their friend to participate in an activity. The functional differences between these tasks will be discussed in further detail in Section 7.4 but for the purposes of general comparison, it is helpful to think of both the letter to the editor and the telephone call tasks as 'argumentative' in the sense that they require linguistic functions that are more similar to the argumentative essay than the picture narrative or oral interview.

Lastly, it is also important to highlight the fact that while all studies investigated the construct of phraseological complexity in terms of diversity and sophistication, they differed in how this construct was operationalized. In the study conducted using the LCF data, phraseological complexity was operationalized as the diversity (RTTR) and sophistication (PMI) of adjectival modifier, direct object and adverbial modifier dependencies. This was done to maximize comparability with Paquot (2018, 2019). However, as mentioned in the discussion section of that chapter, this was not unproblematic. For one, this study raised the question of whether the phraseological units used (in particular adverbial modifiers) were truly representative of the phraseology of French, which as suggested by Vinay & Darbelnet (1995), tends to prefer verbal modification through the use of prepositional phrases (e.g., *répondu avec colère*, 'responded with anger') rather than adverbs.

The LANGSNAP study therefore excluded adverbial modifiers and focused just on adjectival modifiers and direct objects as these seemed to be more representative of the phraseology of L2 French. In addition, rather than RTTR, diversity was operationalized as the number of types over 100-word moving windows. This was also done because in the LANGSNAP data, the RTTR of phraseological units was found to be very highly correlated with text length, a confound that was especially problematic because the oral interviews were much longer in length as compared to the argumentative essays. It also was not possible to use more sophisticated diversity measures such as MTLD because they require a minimum number of tokens (at least 100 as suggested by Koizumi, 2012). One solution could have been to exclude texts that did not have a threshold number of phraseological units but the fact

that some corpus texts contained few to no phraseological units of a given type was seen as an important aspect of register variation (e.g., noun-based units in personal or involved oral tasks). Therefore, the only possible way to calculate diversity was by using a moving average technique. Diversity was calculated as the average number of types over 100-word moving windows (see Section 5.2.2). To ensure comparability between the diversity and sophistication measures used in the LANGSNAP study, sophistication (PMI) was also calculated as the average over moving windows (rather than over entire texts).

In the TEF study, in order to capture phraseological units that are more representative of French (cf. Vinay & Darbelnet, 1995), the decision was made to use four-word lexical bundles rather than adjectival modifiers or direct objects. While this did have the advantage of expanding the type of phraseological units under investigation, it is reasonable to ask whether lexical bundles are indeed comparable with adjectival modifiers and direct objects extracted in the previous two studies. As it turns out, although the lexical bundles are longer, they often encapsulate adjectival modifiers or direct objects within them. To illustrate, Table 7.2 lists the ten most frequent direct object dependencies from the LCF corpus.

Each of these units was searched for within the verb-bundles extracted from the TEF corpus (i.e., those that began with the same verb and contained the noun as one of the remaining three words). Seven out of the ten most frequent direct objects are attested in some fashion within the TEF verb-bundles, which demonstrates that there is some overlap in these units.

This is especially evident in the case of *tenir_VER+compte_NOM* ('to hold account'), which is found as part of the following verb-bundles in the TEF corpus:

- tenir_VER compte_NOM du_PRP point_NOM (PMI 9.95)
  'hold account of the point'
- tenir_VER compte_NOM du_PRP champ_NOM (PMI 9.98)
  'hold account of the scope'
- tenir_VER pas_ADV en_PRP compte_NOM (PMI 4.76)
  'do not hold in account'

- tenir_VER compte_NOM de_PRP ce_PRO (PMI 9.52)
  'hold account of it'
- tenir_VER compte_NOM si_KON ce_PRO (PMI 6.7)
  'hold account if it'
- tenir_VER compte_NOM de_PRP tout_PRO (PMI 9.89)
  'hold account of all'
- tenir_VER compte_NOM de_PRP DET (PMI 9.11)
  'hold account of the'

Table 7.2: Top 10 most frequent direct objects in LCF and correspondence in TEF

	n		FRCOW PMI	
Unit	LCF	TEF	DOBJ	Verb-bundle (range)
commettre_VER+crime_NOM	48	1	8.48	12.21-12.21
tenir_VER+compte_NOM	43	12	6.48	4.76-9.98
prendre_VER+exemple_NOM	38	43	3.32	-1.08-8.67
déguiser_VER+pensée_NOM	35	0	5.69	
trouver_VER+solution_NOM	34	15	4.10	6.08-9.01
rendre_VER+compte_NOM	33	31	5.85	7.84-10.78
violer_VER+loi_NOM	33	0	6.38	
prendre_VER+décision_NOM	29	21	4.52	6.24-9.4
jouer_VER+rôle_NOM	26	31	7.17	7.17-11.5
commettre_VER+délit_NOM	23	0	8.66	

Note:
DOBJ = Direct object

As shown in the list above, *tenir_VER+compte_NOM* is often part of a larger structure that includes a prepositional phrase (e.g., *du_PRP point_NOM*, 'of the point'). This demonstrates that the bounds of phraseological units often go beyond that of a binary direct object relationship. This is not meant to imply that direct objects are not useful. The fact that *tenir* + *compte* has a PMI of 6.48 clearly indicates that the two words have a relatively exclusive relationship. Use of this

phraseological unit therefore demonstrates sophisticated phraseological knowledge on the part of a learner. That being said, widening the scope to include longer units does seem to provide slightly more nuanced information about a learner's phraseological knowledge. One study where this has been done is that of Nesselhauf (2005) who classified verb + object collocations into nine different syntactic categories (e.g., verb + object, verb + preposition + object, verb + object + preposition + object). The results of the analysis showed that learners made more collocational errors on the longer syntactic units (e.g., *answer a question in the affirmative*) compared to 'simple' verb + direct object structures (e.g., *draw a comparison*). Of course, Nesselhauf's study focused on accuracy rather than complexity and used data from L2 English learners but it does speak to the potential usefulness of looking at "other patterns that are less commonly thought of as patterns of verb-noun collocations" (Nesselhauf, 2005: 216). As shown in the list above, not every use of *tenir compte* is equally as 'sophisticated.' This is shown by range of PMI values for each of the verb-bundles containing *tenir compte*, reflecting the strength of the contingency relationship between *tenir* and the rest of the sequence. Focusing solely on the binary direct object relationship masks some of this variation.

Viewed in this way, the lexical bundle approach can be seen as an extension of the methods used in the LCF and LANGSNAP studies, providing a slightly more fine-grained view of phraseological phenomena. It goes without saying that there are also important differences between the methods and not all direct objects or adjectival modifiers will necessarily be captured within the lexical bundles. For instance, any long-distance dependency relationships falling outside of the four-word window will not be captured. Likewise, nouns that are pre-modified rather than post-modified by an adjective (the less frequent variant in French) are not captured by the lexical bundle approach. Still, there is arguably enough similarity between the phraseological units used in the three studies to allow comparison of the results on a general level.

7.2 Phraseological complexity in L2 French (RQ1)

Phraseological complexity aims to capture the size and elaboration of a learner's phraseological repertoire as manifest through the diversity and sophistication of the phraseological units used in production. Although previous research had shown that measures of phraseological complexity were predictive of proficiency (and to some extent also development) in L2 English, studies using other target languages are comparatively rare (cf. Edmonds & Gudmestad, 2021; Rubin et al., 2021; Siyanova-Chanturia & Spina, 2020). The first objective of this thesis was therefore to determine whether Paquot's (2019) construct of phraseological complexity would be a useful index in the context of L2 French. This led to the following research question:

> **RQ1.** To what extent is phraseological complexity an indicator of proficiency and development in L2 French?

The studies using the LCF data (Chapter 4) and TEF data (Chapter 6) addressed the cross-sectional aspect of this research question, comparing the phraseological complexity of texts that were independently rated at different levels of proficiency. The study using the LANGSNAP data (Chapter 5) examined the developmental aspect of this question, tracking the longitudinal change in phraseological complexity within productions of the same group of learners over a period of 21 months. The results of these three studies are summarized in Table 7.3. In this table, upward facing arrows represent cases where a positive relationship was observed between phraseological complexity and proficiency or time. Downward facing arrows represent cases where a negative relationship was observed. All cases where the relationship was found to be non-significant are shown with a line through the arrow. This section will first address the proficiency aspect of this question before moving on to the ability of phraseological complexity to index L2 development.

Table 7.3: Comparison of results across all studies

		Diversity				Sophistication			
		Written	Oral			Written	Oral		
Corpus	Unit	Arg.	Arg.	Nar.	Int.	Arg.	Arg.	Nar.	Int.
Cross-sectional									
LCF	adjectival modifier	↗[1]				↗̸			
	direct object	↗̸				↗[2]			
	adverbial modifier	↗̸				↗̸			
TEF	noun bundle	↗[3]	↘[4]			↗̸	↗̸		
	verb bundle	↗̸	↗̸			↗[5]	↗[5]		
Longitudinal									
LANGSNAP	adjectival modifier	↗̸		↗̸	↘[6]	↗̸		↗̸	↘̸
	direct object	↗̸		↘̸	↘̸	↗[7]		↗[7]	↗[7]

Note:
Arg.: Argumentative task, Nar.: Narrative task, Int: Interview task
Arrows represent the direction of the effect with either proficiency (LCF, TEF) or time (LANGSNAP):
↗: positive effect (sig.), ↗̸: positive effect (n.s)
↘: negative effect (sig.), ↘̸: negative effect (n.s), blank cells indicate cases that were not investigated.

[1] C1-C2 (r = 0.21)
[2] B2-C1 (r = 0.13); B2-C2 (r = 0.33)
[3] 9.36 more points on the exam
[4] 4.69 fewer points on the exam
[5] 7.77 more points on the exam
[6] Decrease in PMI of 0.01/month
[7] Increase in PMI of 0.01/month

7.2.1 Phraseological complexity as an index of proficiency

The LCF data. The first study to compare phraseological complexity across proficiency levels was the study conducted using data from the *Leerdercorpus Frans* (Chapter 4). In this study, phraseological complexity was operationalized as the diversity (RTTR) and sophistication (PMI) of adjectival modifier, direct object and adverbial modifier dependency relations (following Paquot, 2019). These units were extracted from argumentative essays written by university-level learners of French in Belgium. Each essay was independently rated and assigned a CEFR score (B2 - C1).

The results of this study were slightly different from those of Paquot (2019) in that the diversity of adjectival modifiers significantly increased with proficiency from C1-C2 (Paquot found no significant differences across proficiency levels for phraseological diversity). In the discussion of the results of that study, this finding was contextualized however by showing that adjectival modifier diversity was strongly correlated with the diversity of adjectives ($r = 0.89$, $p < 0.001$) as well as the diversity of nouns ($r = 0.75$, $p < 0.001$), which suggested that RTTR may have tapped more into lexical diversity than phraseological diversity.

Although it was not discussed in that chapter, another problem with operationalizing diversity as RTTR is that despite applying Guiraud's (1954) mathematical correction, RTTR still tends to be correlated with the number of tokens (see McCarthy & Jarvis, 2007, 2010). As it turns out, the RTTR of adjectival modifiers is indeed strongly and significantly correlated with the overall number of adjectival modifier tokens ($r = 0.90$, $p < 0.001$). This means that the RTTR value is more reflective of how much adjectival modification a learner used in general, rather than how many *different* adjective + noun combinations were used. Because adjectival modification is an aspect of phrasal elaboration and characteristic of advanced proficiency in general (Biber et al., 2011), it is not surprising that the C1 and C2 level exams differed significantly in this respect. Though not to the same extreme, there are also strong and significant correlations between number of units and the RTTR of

direct objects ($r = 0.89$, $p < 0.001$) and the RTTR of adverbial modifiers ($r = 0.76$, $p < 0.001$). This may explain why no significant differences were found for these measures across proficiency levels. Both structures encapsulate clause-level syntactic features, which are typical of a broader range of discourse and are not *specialized* to academic writing in the same way as phrasal complexity features. If RTTR is simply measuring the quantity of direct objects and adverbial modifiers, it is really just an indirect measure of the number of finite and non-finite clauses and as shown by the random forest model, clausal elaboration was not an important predictor of proficiency. Phrasal elaboration (mean length of noun phases) on the other hand, was an important predictor of proficiency, which is consistent with the significant difference found for the RTTR of adjectival modifiers between C1 and C2 level essays. Because of the overlap between RTTR and other aspects of complexity (i.e., lexical and syntactic) it is not possible to conclusively state whether phraseological diversity was really an index of proficiency in L2 French in the LCF data.

Phraseological sophistication however, was found to be a clear index of proficiency in this data. The mean PMI of direct objects was significantly higher in the C1 and C2 level essays as compared to the B2 level essays and was an important predictor of proficiency in the random forest model. Likewise, although the proportion of non-collocating adverbial modifiers (PMI < 3) was over the threshold for significance (alpha level of 1%) when compared directly across proficiency levels, it was found to be an important predictor in the random forest model. The PMI of adjectival modifiers however was not found to be significantly different across proficiency levels nor was it an important predictor of proficiency in the random forest model. Summing up the results of the LCF study, the general pattern that emerged was that for noun-based phraseological units (adjectival modifiers), the dimension of diversity was more indicative of proficiency than sophistication (notwithstanding the problems with operationalizing it as RTTR) whereas for verb-based phraseological units (direct objects and adverbial modifiers), sophistication was more indicative of proficiency. This general pattern was also evident in the TEF data.

The TEF data. The second study to compare phraseological complex-

ity across proficiency levels used oral and written exam data from the *Test d'évaluation de français* (Chapter 6). In this study, the phraseological units were four-word lexical bundles beginning with either a noun (noun-bundles) or a verb (verb-bundles). Diversity was operationalized as the ratio of bundle types to POS-structures and sophistication as the PMI of the first word and the rest of the sequence. In order to control for text length, both measures were calculated as the average over ten random samples of five units. Just as in the LCF data, for noun-based units, diversity was more predictive of proficiency than sophistication.[2] In the case of writing, this was a positive effect. Holding all else constant, a one standard deviation increase in noun-bundle diversity was associated with 9.36 more points on the written portion of the exam. For the oral data, a one standard deviation increase in noun-bundle diversity was associated with 4.69 *fewer* points on the exam. A more detailed discussion of cross-modal differences will follow in Section 7.4 but here it will suffice to say that although the direction of the effect was different for written versus oral production, in both cases, sophistication seems to have had a stronger effect on proficiency scores than diversity (which was positive but n.s. across both modes). The opposite pattern was found for verb-bundles. Although diversity was not found to be significantly correlated with proficiency, a one standard deviation increase in the PMI of verb-bundles was associated with 7.77 more points on the exam (difference between oral and written n.s.).

The results of the LCF and TEF studies are generally in line with previous research in other languages. For one, these results provide more evidence that diversity seems to be more indicative of lower levels of proficiency. As was the case of the TEF data, all studies that have found significant proficiency differences in phraseological diversity have included learners below the B2 level (Garner, 2020;

[2] Note that just because there were convergent results for 'diversity' in both the LCF and TEF studies, it does not mean that they were necessarily measuring the same thing or that the operationalization of diversity as RTTR in the LCF study was unproblematic. This shown by the fact that the correlation between the number of units in a text and the random sample-based diversity measures is low for both noun-bundles (r = -0.08) and verb-bundles (r = -0.16). This suggests that these measures were somewhat better able to tap into the dimension of variation (i.e., the number of different types) rather than volume (i.e., the total number of types).

Jiang et al., 2021; Paquot et al., in press; Rubin et al., 2021). The data presented here also show that phraseological sophistication measures are indicative of proficiency in L2 French, which is consistent with previous research comparing the PMI of phraseological units across proficiency levels in written production (Bestgen, 2017; Garner et al., 2018, 2019; Granger & Bestgen, 2014; Jiang et al., 2021; Kim et al., 2018; Paquot, 2018, 2019; Paquot, Naets, et al., 2021; Rubin et al., 2021; Zhang & Li, 2021) and oral production (Eguchi & Kyle, 2020; Uchihara et al., 2021; Zhang et al., 2021). In contrast to studies that have found a negative relationship between PMI and oral proficiency (e.g., Paquot et al., in press; Tavakoli & Uchihara, 2020), no such relationship was found in the cross-sectional TEF data (but see Section 7.2.2 for negative relationships found in longitudinal data). Putting everything together, the results of the LCF and TEF studies show that phraseological complexity measures can be useful predictors of proficiency in L2 French. That being said, one thing that is evident is that the two dimensions of diversity and sophistication are not necessarily equally as informative about all types of phraseological units in L2 French or about all levels of proficiency. At least in the two studies presented above, diversity seems to be more indicative of proficiency for noun-based units (though the direction of the effect differs across modes) and sophistication seems to be more indicative of proficiency for verb-based units.

It is also important to note that even when there was a significant relationship between phraseological complexity and proficiency, this was often a very small effect. In the LCF data, for example, the effect sizes (r) for significant pairwise differences in proficiency levels ranged from 0.13 to 0.33. Likewise, In the TEF study, a one standard deviation increase in phraseological complexity (diversity or sophistication) amounted to fewer than ten extra points on the exam. This shows that while phraseological complexity appears to factor into raters' perceptions of a learner's proficiency, other factors are also quite important. On the face of it, this should not be surprising given that scoring rubrics ask raters to take many aspects of linguistic competence (e.g., grammar, lexis, accuracy) as well as communicative ability (e.g., task completion) into account. Some of these factors may actually in-

teract with phraseological complexity measures, mediating their on proficiency ratings.

One such factor may be the influence of morphosyntactic deviances (e.g., non target-like gender or number agreement in attributive adjectives), which has been shown to persist even at very advanced stages of L2 French, both in written (Forsberg & Bartning, 2010) as well as oral production (Bartning et al., 2009). As suggested by Stengers et al. (2011), because inflectional errors are salient to raters, their presence may dampen the contribution of phraseology to holistic evaluations of proficiency. Practically this means that if, for example, a learner uses a phraseological unit with a very high PMI but non-targetlike gender or number agreement, the overall impression may be more negative than positive, resulting in a lower proficiency score (for a text with high phraseological complexity). Interestingly, when comparing the number of morphosyntactic deviances across proficiency levels, Forsberg and Bartning (2010) found that there was a higher percentage of errors in noun phrases than in verb phrases. This might explain why in both the LCF and TEF data, the sophistication of noun-based phraseological units was not significantly associated with proficiency. If noun phrases contain more errors overall, this could have contributed to lower scores despite high levels of phraseological sophistication. This is illustrated in Example 7.1. This extract is from an exam with the second highest average noun-bundle PMI in the TEF corpus (9.65). Yet, despite the use of relatively sophisticated bundles (e.g., *viande_NOM et_KON produit_NOM laitier|laitière_ADJ*, 'meat and dairy products,' PMI = 12.99), this exam received a score of only 395 (CEFR B1). Looking at the text in closer detail shows that there are several non-targetlike forms which may have contributed to the lower score (e.g., *de les viandes*). In the noun-bundle itself, not only is the word *latiere* ('dairy') missing an accent (i.e., *latière*), but the adjective also does not agree in number or gender with the masculine and plural noun *produits* ('products').

Example 7.1. Extract from a low scoring text with a high average noun-bundle PMI (839715-EE16086-W-Ba.txt):

> "Cependant, c'est possible d'aggraver la santé en refusant

de les viandes et produits-M.PL laitiere-F.SG"

However, it's possible to compromise your health by refusing meat and dairy product.

In order to make the oral and written sections of the exam comparable, all spelling errors were corrected and lemmatized which means that spelling and inflectional errors such as these did not have an effect on phraseological complexity measures even though they were likely very noticeable to raters. This example exemplifies a potential trade-off between complexity and accuracy, which is well-documented in the TLBT literature. As argued by Skehan, (1998: 179), this manifests as a competition between "conservatism and correctness (accuracy) and risk-taking and change (complexity)." Given that the focus of this thesis is on complexity, a more detailed comparison of the relationship between phraseological complexity and accuracy is somewhat beyond the scope here but these findings suggest that it would be a worthy area for future research.

7.2.2 Phraseological complexity as an index of development

The LANGSNAP Data. The development of phraseological complexity over time was investigated using the LANGSNAP data (Chapter 5). In this study, phraseological complexity was operationalized as the diversity (number of types) and sophistication (PMI) of adjectival modifiers and direct objects (both averaged over 100-word windows), which were extracted with the use of regular expression searches from argumentative essays, picture-based oral narratives and oral interviews conducted over a period of 21 months from university-level learners of French. As discussed in Section 7.1, it is difficult to directly map Elicited Imitation Test (EIT) scores to the standardized CEFR scores used in the LCF and TEF studies but given that the learners were studying French at the university level and had, on average, 10.5 years of previous French education, it is reasonable to assume that they were relatively advanced. In terms of general linguistic development, Mitchell, Tracy-Ventura and McManus (2017: 79) previously reported that the average general proficiency of the learners

(as measured by EIT scores) increased from 62.9/120 to 80.1/120 at the second collection point eight months later, a change that was found to be significant ($p < 0.001$) with a medium effect size ($d = 0.99$). At postsojourn (15 months into the study), the average EIT score was 85.6/120, an increase that was also significant ($p < 0.001$) and equated to a small effect size ($d = 0.34$).

In light of the fact that the learners had appeared to increase in general proficiency, at least as measured by EIT scores, it was also reasonable to assume that changes would also be seen over time in the phraseologial complexity of their productions. In terms of phraseological diversity, although positive trends were observed for several measures, the only significant change over the study period was for the diversity of adjectival modifiers, which actually decreased (significantly) by 0.01 per month in the oral interviews. In the discussion of that chapter, this was explained by the fact that learners had most likely already reached a level of phraseological diversity that was necessary to complete those tasks. This was shown by the fact that the estimated marginal mean for these university-level learners at both the beginning and end time points was generally within the confidence intervals for the L1 group on each of the tasks.

An important takeaway from this study was therefore that in order to observe development, there must be scope for development. If tasks themselves do not promote high levels of phraseological diversity, then there is little reason to expect to observe it in production given that learners respond to a task's "semantic and pragmatic demands" (Pallotti, 2009: 599). It is also important to keep in mind that the study abroad context may not have lent itself to the development of phraseological complexity. This may be because the majority of encounters that learners had with native speakers were probably quite informal (e.g., conversations with colleagues or with fellow members of social clubs) and as shown by Arvidsson (2019), amount of target-language contact during study abroad is not necessarily a significant predictor of the phraseological knowledge upon return home. That being said, the LANGSNAP corpus does actually contain extensive information regarding the linguistic experiences of participants collected through questionnaires. While it was not possible within the scope

of this project to correlate that data with measures of phraseological complexity, it may be a fruitful area for future research.

It may also be the case that increases in diversity are only visible over a much longer time frame. The only longitudinal study to show a significant increase in diversity over time is that of Qi & Ding (2011), who reported a significant increase in the average TTR of manually-identified 'formulaic sequences' over four years of university education in the oral monologues of Chinese learners of English. No proficiency information about the learners was provided by the authors, but these were university-level learners of English so it is reasonable assume that they were at a similar proficiency level to the LANGSNAP participants. What is striking in these results is that even though the authors found a significant increase in diversity, this only amounted to 3.63 more types for every token, or about 0.06 new types per month, which again points to the fact that increases in phraseological diversity are likely very gradual and likely only observable over extremely long time periods.

In terms of sophistication, on average, the LANGSNAP learners started out lower than the L1 group and the PMI of both adjectival modifiers and direct objects increased slightly over the course of the study period. However, this increase was only found to be significant in the case of direct objects , which increased by 0.01 per month (task differences n.s.) and not adjectival modfiers (in contrast to the results of Edmonds & Gudmestad, 2021). It is notable that it was the sophistication of the verb-based phraseological unit that showed significant development, echoing the cross-sectional results of the LCF and TEF studies. This suggests that not only is direct object sophistication salient to proficiency raters, but it also appears to be (slightly) more susceptible to development than the sophistication of adjectival modifiers. Specifically, the obligatory nature of direct objects may make them more noticeable to learners than adverbial modifiers. In the discussion of Chapter 5, an example was cited where a learner seemed to have struggled to produce certain direct objects when initially completing the oral narrative task prior departing for France but was able to produce them a year later when completing the same task again. Extracts from those two productions are provided in Example 7.2 and

7.3 respectively. What is evident in the two examples is that the learner is clearly able to use much more specific direct objects following her time spent in France, including not only *lire + conte* ('read fairy tale,' PMI = 2.89) but also *chasser + papillon* ('hunt + butterfly,' PMI = 4.37). To be sure, the second extract does include an adjectival modifier that was not initially present (*travail + manuel*, 'manual+labour/crafts,' PMI = 3.44) and this does provide more detail to the story. However, the most important events are conveyed through direct objects and using more sophisticated units allows the learner to more precisely narrate the events of the story (e.g., the cat did not just 'play in the garden,' he hunted butterflies).

Example 7.2. Cat story (C111, May 2011)

> "alors l'histoire de Natalie et de son chat Pompon. tous les matins étaient pareils. chaque jour Natalie aime bien de lire à ses de peut être. elle aime bien de **fait l'art** et de **créer les maisons** de les boîtes. elle aime de jouer avec ses amis dans les vélos. et elle aime de fait de l'exercice dans sa jardin. Pompon aime de dormir dans sa lit. quelquefois il sort. et il aime bien de jouer dans le jardin dans les arbres et avec les papillons. et il aime bien de baigner dans le soleil."
>
> *so the story of Natalie and her cat Pompon. every morning was the same. every day Natalie really likes read to her to maybe. she really likes to* ***make art*** *and to* ***create houses*** *out of boxes. she likes to play with her friends in the bikes. and she likes [to] do exercise in the garden. Pompon likes to sleep in her bed. sometimes he goes out. and he really likes to play in the garden in the trees and with the butterflies. and he likes to bathe in the sun.*

Example 7.3. Cat story (C111, May 2012)

> "il était une fois qu'il y avait une fille s'appelle Natalie et un chat Pompon. Pompon était le chat de Natalie. tous les matins étaient pareils. Natalie **lisait les contes** au ses petits poupées. et elle elle **faisait les arts** les travaux manuels. elle jouait avec ses amis. et elle jouait aux aires de jeux avec sa petit ours. Pompon lui il dormait. et il sortait de la mai-

> son. et il jouait au jardin. il aimait beaucoup de **chasser les papillons** et de baigner sous la soleil."
>
> *once upon a time there was a girl called Natalie and a cat Pompon. Pompon was Natalie's cat. every morning was the same. Natalie **read fairy tales** to her little dolls. and she she **did arts** and crafts. she played with her friends. and she played in the playground with her little bear. Pompon slept. he went out of the house. and he played in the garden. he liked to **hunt butterflies** and to bathe under the sun.*

The learner cited above was surely aware that her vocabulary was lacking when she attempted (and then failed) to produce *lire + conte* as shown by the fact that she stopped and had to reformulate the sentence, eventually abandoning it altogether. It is possible that this experience made the unit more salient to her when she encountered it later. No such hesitation or reformulation phenomena are observed in the section where the adverbial modifier *traveaux manuels* later emerged however, due perhaps to the fact that she was able to tell the story perfectly well without that phraseological unit. Of course, this is pure speculation and there is a limit to what can be said on the basis of this one example but it does seem to be illustrative of the differences between the two types of phraseological units. Considered together with previous research showing that verb + noun collocations are particularly challenging for learners (e.g., Altenberg & Granger, 2001; Nesselhauf, 2005), this seems to suggest though that direct objects, and perhaps verb-based phraseological units more generally, may be more indicative of development relative to adjectival modifiers and other noun-based phraseological units.

What is also notable is that although the rate of development did differ slightly between tasks, this difference was not significant, which may suggest that direct object sophistication is a more reliable measure of development across different tasks. As discussed in Section 6.5, this may be because verbs and direct objects are required for a variety of different communicative functions. This contrasts with adjectival modifiers, which as a phrasal-level feature, tend to be more

common in information heavy (written) registers. In general, the longitudinal results of the LANGSNAP study are also in line with previous research that have found either no significant growth (Bestgen & Granger, 2014; Garner & Crossley, 2018; Kim et al., 2018; Yoon, 2016) or very slow patterns of development (Edmonds & Gudmestad, 2021). As argued by N. C. Ellis et al. (2008), knowledge of highly exclusive word combinations (i.e., those with a high PMI) requires an extraordinary amount of input, not just of the individual word combinations but of the relative frequencies of each word in combination with other words, knowledge that surely takes a very long time to develop. This development is therefore only likely to be seen on a much larger timescale and then only when tasks allow learners to showcase development, which explains why the effects of phraseological complexity are so much weaker in longitudinal data than in cross-sectional data.

7.3 Phraseological complexity in relation to other domains of complexity (RQ2)

The second objective of this thesis was to determine how phraseological complexity measures differ from other measures of complexity that are typically used in L2 research. In Paquot's (2018, 2019) studies of L2 English, she found that measures of phraseological complexity were better predictors of proficiency than measures of lexical or syntactic complexity from B2 to C2. Given the synthetic nature of French, morphological complexity in particular was hypothesized to play more of a role in L2 French (in line with the claims of Stengers et al., 2011). This led to the following research question:

> **RQ2.** To what extent does phraseological complexity relate to other aspects of linguistic complexity as tapped by the current repertoire of measures of syntactic, lexical and morphological complexity in L2 French?

This question was primarily investigated in the partial replication study of Paquot (2018, 2019). In order to replicate this study as closely as possible, it was necessary to first develop an automatic tool for the calculation of syntactic complexity measures. This section therefore

begins with a discussion of the *fsca* tool described in Chapter 3 before moving on to the relationship between phraseological complexity measures and other domains of complexity. Within the scope of this thesis, it was not possible to compare all types of complexity across modes so this discussion is primarily focused on the results of the LCF study (Chapter 4) but reference is also made to the results of the TEF study (Chapter 6), which included lexical complexity measures.

7.3.1 The fsca tool

The development of the *fsca* tool was necessitated by the fact that no such automatic tool for the calculation of syntactic complexity measures previously existed for L2 French. This means that studies of syntactic complexity in L2 French have typically been limited in terms of the numbers of syntactic measures they have been able to calculate, given the time-consuming and costly nature of manual annotation. The development of the *fsca* tool revealed that manual annotation of these units is also highly subjective. Even syntactic constituents that one would think are relatively straightforward to identify (e.g., coordinated clauses) had low levels of interrater agreement. This is all the more troubling given the fact that few studies report reliability statistics for syntactic annotation. The comparison of the manual annotation with the output of the automatic tool also revealed that the correct identification of syntactic units largely depends on the accuracy of pre-processing steps (e.g., segmentation, POS-tagging, dependency parsing), each of which can introduce errors.

Despite these errors, all syntactic measures calculated with the *fsca* tool were found to be strongly and significantly correlated with manually-derived measures, which was taken as sufficient justification to use them in the partial replication study. However, this does not mean that the use of these measures was unproblematic. For one, the strength of the correlations between manual and automatic measures varied greatly which certainly introduced a level of error into the analysis of the LCF data in Chapter 4. Notably, the results of the random forest showed that only one measure of syntactic complexity (mean length of noun phrase) was an important predictor of

proficiency. It is theoretically possible that other measures of syntactic complexity may also have been predictive of proficiency had their extraction been more reliable. Even in the case of MNLP, the strength of the correlation between the manual- and automatically-derived version of this measure (ρ = 0.720) was below Plonsky and Derrick's (2016) recommended level for interrater reliability for L2 research (there are no known guidelines for the reliability of automatic annotation). Error analysis revealed that this measure is highly affected by incorrect dependency annotation, for example, analyzing a verbal complement as being dependent on the immediately preceding noun rather than the verb that heads it. This means that at least some of the variation in proficiency scores attributed to "noun phrase length" by the random forest model may actually be caused by clausal complexity outside the noun phrase itself. Another problem with this measure is that as an 'omnibus' measure of complexity, MLNP collapses relevant structural and functional differences together (Biber et al., 2011, 2020, 2021). A noun phrase can be made longer in a number of different ways (Norris & Ortega, 2009), including through the use of embedded clauses, which means that MLNP can represent both phrasal complexification (e.g., nominalizations, adjectival modification) as well as clausal complexification (e.g., addition of relative clauses). Unfortunately it was not feasible to develop and test more specific complexity measures in the context of this project so the decision was made to use the same measures as Paquot (2019) to ensure maximum comparability. That being said, there is clearly a need to develop tools for L2 French that can extract a broader ranger of different grammatical structures (e.g., non-clausal phrases) and that can distinguish between the functions of such structures to enable a more "linguistically interpretable" description of learner behaviour (to borrow the words of Biber et al., 2020, 2021).

7.3.2 Lexical complexity

In the partial replication of Paquot (2018, 2019), lexical diversity was operationalized using type-token ratios with Guiraud's (1954) transformation applied (separate measures were calculated for all words, all lexical words together as well as nouns, verbs, adjectives and

adverbs separately). In addition, following Paquot, ratios of each part of speech to the total number of lexical words were also calculated (e.g., noun types/all lexical tokens). Lexical sophistication was operationalized as proportions words absent from the French Frequency Dictionary (Lonsdale & Le Bras, 2009) and the proportion of words present on the Lexique Transdisciplinaire list (Tutin & Grossman, 2014), again with Guiraud's transformation applied and divided into both general measures representing all words and separate measures for each part of speech. In the random forest model, two measures of lexical diversity (noun variation and lexical word variation) and two measures of lexical sophistication (lexical words absent from the FFD and nouns absent from the FFD) were found to be important predictors of the holistic CEFR scores assigned to texts by the raters. The partial dependency plots revealed that there seemed to be a trade-off between lexical diversity and sophistication with proficiency. B2 texts tended to exhibit less lexical sophistication but relatively high levels of lexical diversity whereas C1 texts were less lexically diverse than B2 texts but had, in general, higher levels of lexical sophistication. Only at the C2 level did texts have both high levels of lexical diversity and sophistication. Taken together, this seemed to indicate that raters were sensitive to the lexical complexity of the essays and it contributed to their classification into CEFR levels.

This finding was somewhat contradicted by the results of the TEF study however where lexical diversity and sophistication seemed to be less-useful indicators of proficiency scores. In that study, lexical diversity was operationalized as MTLD[3] and lexical sophistication as the number of words absent from the top 2000 most frequent lemmas in the FRCOW16[4] corpus (averaged over 100-word moving windows). In the TEF data, which had a larger range of proficiency scores (CEFR A1-C2), while lexical diversity did contribute to proficiency scores, this relationship plateaued such that an increase greater than 1 standard deviation in MTLD no longer contributed to additional points on the

[3]This measure found to be less strongly correlated with text length than other TTR-based measures or transformations (r = -0.54).

[4]This measure was used so that lexical sophistication would be based on the same reference corpus as the phraseological sophistication measures to maintain comparability between the two.

exam. For lexical sophistication, while there was a linear correlation with proficiency scores, this was only for the minority of texts on the topic of travel.

Comparing the LCF and TEF studies, one major difference is the homogeneity of the learner samples. The LCF data is rather homogeneous and contains data only from university-level learners of French in Belgium whereas the TEF corpus contains data from self-identified L1 and L2 users of French at a wider range of proficiency levels (A1-C2).[5] Ostensibly, this could be the reason for the conflicting findings. While lexical complexity measures have been relatively useful in distinguishing between university-level learners of French (as in Tidball & Treffers-Daller, 2007, 2008), they have not been able to capture differences between very advanced learners and native speakers (as shown by Forsberg Lundell & Lindqvist, 2014; Lindqvist et al., 2011). In contrast, phraseological complexity measures were useful predictors of proficiency in both the LCF and TEF data, which seems to speak to their general usefulness as indices of advanced proficiency although it is, of course, important to keep in mind that different types of phraseological measures may be better than others at distinguishing between these advanced learners (as discussed in Section 7.2).

7.3.3 Syntactic and morphological complexity

In the LCF study, syntactic complexity was operationalized using length measures, diversity measures (standard deviation of length measures) and ratio measures at the levels of the sentence, T-unit, clause and noun phrase. Morphological complexity was operationalized using the Morphological Complexity Index (Pallotti, 2015) which counts the average number of inflectional variants of verbs (e.g., m=ind|n=s|p=3|t=pst, indicative singular third person present tense) that were used in a given text. Only one syntactic complexity measure was found to be an important predictor of CEFR level in the random forest model, namely the mean length of noun phrases. The partial dependency plots revealed that this measure mostly distinguished between B2 and C1 level texts. Texts with noun phrases that were

[5] Note that while there were texts covering the whole range, there was a skew towards the upper levels as shown in Table 7.1.

greater than 3.41 words on average were more likely to be classified at the C1 level than B2 level. Morphological diversity was also found to be an important predictor of proficiency in the random forest model, distinguishing primarily between B2 texts, which used up to 5.46 different verbal inflections (on average across random samples of ten verbs) and C2 level texts, which used up to 8.66 different verbal inflections. Notwithstanding the problems associated with MLNP as described in Section 7.3.1, this speaks to the importance of both syntactic and morphological complexity to evaluations of proficiency in L2 French. In general, this finding is in line with proposed *developmental stages* for L2 French (Ågren et al., 2012; Bartning & Schlyter, 2004) which claim that the advanced stages of L2 French are characterized by the use of a variety of morphological inflections as well as "the integration of clauses characterizing a capacity to handle many levels of information within the same utterance" (Bartning & Schlyter, 2004: 296).

At the same time, there appears to be a limit to how much morphological phenomena contribute to differences between very highly advanced levels of proficiency. For example, Forsberg Lundell et al. (2014) found that the proportion of morphosyntactic deviances in learner texts could not distinguish between highly advanced speakers (long-term residents of France) and native speakers. In comparison to the data used by Forsberg Lundell et al., the LCF corpus, is comprised of data from university-level learners who may not have quite reached the same proficiency level. This may explain why according to the random forest model, morphological diversity was actually more important to predictions of proficiency than the phraseological complexity measures. Even for these learners though, no syntactic complexity measure contributed to differences between C1 and C2 level texts.[6] Taken together, these results showed that, although phraseological complexity measures were predictive of proficiency, so too were lexical, syntactic (phrasal) and morphological measures (in contrast to Paquot's 2018, 2019 results for L2 English).

[6]Perhaps also because the omnibus syntactic measures used were not sensitive enough to pick up on the specific grammatical phenomena that are indicative of advanced writing (cf. Biber et al., 2011, 2020, 2021).

7.4 Phraseological complexity across modes (RQ3)

The final objective of this thesis was to compare phraseological complexity measures between written and spoken production. Cross-modal differences were hypothesized to be due to two primary factors: the amount of attention learners can pay to output in the two modes (Kormos, 2014; Skehan, 2014) and the differences in communicative functions between oral and written tasks (Biber, 1988; Biber et al., 1999). This led to the following research question:

> **RQ3.** To what extent does phraseological complexity differ in oral versus written L2 production?

This question was addressed via the longitudinal LANGSNAP study presented in Chapter 5 as well as the cross-sectional TEF study presented in Chapter 6. The following section will explore how both production-related factors as well as functional factors appear to have contributed to cross-modal differences in phraseological complexity.

7.4.1 Production-related factors

In Chapter 2, one of the principle reasons for cross-modal differences was hypothesized to be due to production-related factors. Specifically, it was argued, in line with the TBLT literature, that writing eases the burden on the formulator leading to more complex output in written as opposed to spoken L2 production.

This was indirectly tested in the LANGSNAP study by comparing L1 and L2 performance on the three tasks. The assumption was that language production of the L1 group is likely more automatized, which means that there should be less 'burden' on the formulator compared to the learners. As a result, the formulator should behave relatively similarly across both modes and thus any differences observed between tasks should be attributable to functional, communicative factors rather than processing factors. In other words, the assumption was that, putting register aside, written production would not be *inherently* more complex for L1 users because the formulator should be

able to meet the demands of the conceptualizer in both modes (or at least to a greater extent than is possible for learners).[7]

As it turned out, this indeed seemed to be the case. Comparing across tasks, there was only one measure of phraseological complexity (diversity of adjectival modifiers) that was systematically higher in the argumentative essay as compared to both the oral narrative and oral interview for the L1 group. For the three other measures of phraseological complexity, the values were not significantly different between the essay and the narrative but significant differences were found between both the essay and the narrative on one hand and the interview on the other. Put differently, the written essay and the oral narrative were more similar to each other than they were to the interview. This seemed to suggest that the L1 group was not constrained in the selection of phraseological units by the real time processing demands of the oral mode and that task differences were more likely to be caused by register-related factors (i.e., the selection of specific units for a communicative purpose). The use of the L1 group served as a 'baseline' for comparison to the learner group. If it could be shown that the learners behaved differently across two tasks where the L1 group behaved similarly, it could be taken as evidence that such differences were due to processing rather than register related factors. Except for one measure (direct object sophistication), the learner group indeed had significantly higher levels of phraseological complexity in written tasks as compared to oral tasks. In the discussion of that chapter, it was argued that writing enabled learners to decrease the burden on the formulator, helping them to use a greater variety of phraseological units and to use phraseological units that were more sophisticated.

In the TEF study, the focus was not on the comparison of the different tasks directly, but rather on the effect phraseological complexity had on oral versus written proficiency scores. In general, it was found that although the effect was not always strong (or even significant), phraseological complexity tended to be positively correlated with proficiency

[7] To be clear, this is not meant to suggest that writing and speaking exhibit the same types of complexity features. Quite to the contrary in fact, as has been argued throughout this thesis (and will be discussed further in Section 7.4.2). Rather, the point here it that it is important to consider production-related differences as well as register-related differences when comparing L2 speech and writing.

in both oral and written production. One exception to this was the diversity of noun-bundles, which was negatively correlated with oral proficiency scores. The explanation given for this unexpected finding was that the oral mode may have promoted the use of a small set of semi-fixed units, in line with Leech's claim (2000: 697) that speech relies on a "restricted and repetitive lexicogrammatical repertoire" due to the constraints of real-time processing. It was found that noun-bundles were repeated with only a one-word difference 0.55 (SD = 0.94) times on average in oral productions, compared to 0.18 (SD = 0.45) times in written productions, leading to small differences such as *temps avec tes enfants* ('time with your children'; PMI = 4.5) and *sport avec tes enfants* ('sport with your children'; PMI = 3.05) within the same utterance, which may have been negatively perceived by raters.

Of course, there is a limit to how much can be inferred about a learner's cognition on the basis of corpus data alone but the results of the LANGSNAP and TEF studies do seem to indicate that processing-related factors have an effect on the phraseological complexity of L2 production. As discussed in the following section though, this effect was often mediated by the communicative function of the task.

7.4.2 Register-related factors

As shown by register-based approaches to linguistic variation, communicative function goes hand in hand with the selection of linguistic features (Biber & Conrad, 2019). Specifically, corpus research has consistently found that oral and written registers in different languages tend to be distinguished by an opposition between a personal/involved focus and an informational focus (Biber, 2014). Linguistically, this is manifest through the use of clausal features in oral registers and phrasal features in written registers. It was argued in Chapter 2 that such patterns should also extend to the level of phraseology to the extent that phraseological units overlap with phrasal versus clausal syntactic features. Broadly speaking, the phraseological units used in the empirical studies of this thesis can be classified into noun-based units and verb-based units as shown in Table 7.4.

Table 7.4: Classification of phraseological units used across studies

Noun-based units	Verb-based units
adjectival modifier	direct object
noun-bundle	adverbial modifier
	verb-bundle

The noun-based units tend to target complexity at the phrase level. Adjectival modifiers, for example, are located within noun phrases and contribute to phrasal complexity by compressing information through nominalization (e.g., *déliquance + juvénile*, 'young delinquency'). In a similar way, noun-bundles often tap into phrasal complexity because they capture post-modification of nouns (e.g., *salutation_NOM DET plus_ADV distinugué_ADJ*; 'most distinguished salutation'; *expérience_NOM de_PRP DET vie_NOM*, 'experience of DET life'). Of course, because noun-bundles are based on surface-level POS tags, they do not always represent phrasal elaboration. For example, one of the most frequent noun-bundles in the TEF corpus is *part_NOM de_PRP DET avis_NOM*, which is part of the larger expression *faire part de (mon) avis* ('share (my) point of view'). In this expression, the prepositional phrase *de mon avis* is a verbal complement rather than a modifier of the noun *part* and so looking at the surface level POS tags masks the deeper syntactic structure. Nevertheless, as a generalization, noun-bundles can be said to be more indicative of phrase-level complexity than clause-level complexity. On the other hand, direct objects, adverbial modifiers and verb-bundles are each centred around verbs rather than nouns and therefore are more likely to represent clausal complexity features. The initial hypothesis posited in Chapter 2 was that these differences would mean that the complexity of noun-based phraseological units would be higher in written, informational tasks whereas the complexity of verb-based phraseological units would be higher in oral/involved tasks.

As it turned out, this was somewhat of an over-generalization. While tasks may differ in the extent that they rely on *phrasal* complexity

Table 7.5: Phraseological unit tokens in LANGSNAP data

	Median 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)

Table 7.6: Phraseological unit tokens in TEF data

	Median per 100 words (IQR)	
Task	Noun-bundles	Verb-bundles
Letter to the Editor	7.5 (2.75)	6.58 (2.12)
Telephone Call	3.42 (1.23)	6.25 (1.76)

Note:
Counts of noun- and verb- bundles in this table include only those ocurring at least 5 times in the reference corpus.

features (and thus noun-based phraseological units), clausal syntactic features, and by extension verb-based phraseological units, are required by all tasks. This is illustrated in Tables 7.5 and 7.6. It is true that there are, on average, fewer noun-based units in oral tasks relative to the amount found in written tasks. However, within each task individually, there are more verb-based units than noun-based units (with the exception of the 'letter to the editor' task).

This may explain why it was the diversity of adjectival modifiers but not direct objects that was found to differ significantly between L1 written and oral tasks in the LANGSNAP data. While direct objects were common to all tasks, adjectival modifiers served more of a functional role in the essays than in either the narratives or interviews because they allowed writers to densely encode information relevant to the argument they wanted to make. This was illustrated in Example 5.1 (recopied below) where the writer's use of the long noun phrase (beginning with "une conséquence...") allowed her to make a more nuanced argument by balancing several different propositions:

Example 5.1

> "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*.

This dense packaging of information was less important in both of the oral tasks, where the goal of the speakers was to communicate the events of the picture story or their lives (see Examples 5.2 and 5.3, also recopied below)

Example 5.2

> "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*.

Example 5.3

> "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*.

To be sure, it is not that direct objects were not useful in argumentative essays, but rather that what made the argumentative essays

distinctive was their use of adjectival modifiers.

When it came to proficiency ratings of learner texts, raters also seemed to be more sensitive to the diversity of noun-based phraseological units than verb-based units. For example, in the LCF data, there was a significant difference in adjectival modifier diversity[8] between C1 and C2 level essays but no such difference for the direct object diversity or adverbial modifier diversity. In the random forest analysis, this measure did not prove to be an important predictor of proficiency scores but this measure was highly correlated with the diversity and sophistication of lexical words (including adjectives) and nouns, measures that *were* important predictors of proficiency scores.

In the TEF data, the diversity of noun-bundles was also found to be significantly correlated with written proficiency scores. High scoring texts tended to be informationally dense and made use of long noun phrases with extensive post-nominal modification. An extract from one such text is provided in Example 7.4.

Example 7.4. Extract from a high-scoring TEF written exam (899945-EE16083-W-Ba.txt, CEFR C2):

> "...Je découvre avec surprise votre article sur l'attractivité touristique mondiale dans lequel vous citez une certaine normalisation de l'activité, quelque soit la destination. Je ne pense pas que ce soit le tourisme qui fasse l'identité d'un pays ; nous voyageons pour découvrir de nouvelles cultures, pour connaitre l'histoire d'un pays..."
>
> *...I discover with surprise your article about the attraction of world travel in which you cite a certain normalization of the activity, no matter the destination. I do not think that it is tourism that makes the identity of a country; we travel to discover new cultures, to learn about the history of a country...*

In the extract, this candidate uses three different noun-bundles in the

[8] This measure (RTTR) was highly correlated with the number of adjectival modifiers overall (see Section 7.2.1) so this means that the higher rated essays essentially just used more adjectival modification overall.

structure NOM PRP DET NOM : *normalisation de l'activité* ('normalization of the activity'; PMI = 5.55), *identité d'un pays* ('identity of a country'; PMI = 5.65), *histoire d'un pays* (history of a country; PMI = 6.09). Just as in the LCF and LANGSNAP data, the use of noun-based phraseological units appears to contribute to the ability to densely encode information relevant to the argument that the writer is attempting to make.

The TEF data revealed, however, that there was a negative relationship between noun-bundle diversity and oral proficiency scores. As discussed in the preceding section, this seems to be because the pressure of real-time production promoted the use and re-use of a small set of slightly modified noun-bundles (i.e., those that differed by only one or two words) in low-scoring exams. High scoring oral productions, on the other hand, tended to be much more judicious in their use of noun phrases. This is illustrated by the extract in Example 7.5.

Example 7.5. Extract from a high-scoring TEF oral exam (878283-O-B.txt, CEFR C2):

> "Donc en plus je sais que ça c'est important pour toi d'amener ta petite pierre à l'édifice justement par rapport à la planète pour manger mieux pour moins utiliser ta voiture etc là justement c'est voilà ça concilie les deux à la fois...Je pense que tu peux du coup doubler ton salaire avec les pourboires que tu pourras avoir..."
>
> *So also I know that it's important for you to do your bit when it comes to the planet to eat better to use your car less etc so exactly it's that take care of the two things...I think that you can even double your salary with the tips you could make...*

To be sure, this extract does contain noun-bundles, for example, the relatively sophisticated *pierre à l'édifice* (PMI = 14.62), from the idiom *amener/apporter sa pierre à l'édifice* ('to do one's part') but the goal of the task (convincing a friend to participate in an activity) seems to have promoted the use of more personal or involved style of discourse as shown through the frequent use of personal pronouns and subordinated clauses rather than long noun phrases.

The fact that the diversity of noun- rather than verb-based units is associated with proficiency is in line with the claims of Biber et al. (2011). Compared to phrase-level syntactic complexity features, clausal complexity features, are used across a wide range of different communicative settings. This means that a text with a variety of different direct objects, adverbial modifiers or verb-bundles is not necessarily indicative of proficiency in the same way that a text with heavy phrasal modification may be. Because they are distinctive of academic discourse, phrasal complexity features tend to be acquired later (Biber et al., 2011), which may make the use of such features, and by extension noun-based phraseological units, more marked to proficiency raters across a variety of different oral and written tasks (as illustrated by Figure 2.6 in Chapter 2). At the same time, these results also expand on those of Biber et al. (2011) because they show that what marks highly proficient texts is not only the use of phrasal structures (i.e., quantity) but more specifically the use of a variety of different word combinations within those structures (i.e., diversity).

The fact that verb-based phraseological units were used across all tasks also explains why sophistication rather than diversity of those units tended to be correlated with proficiency. If all tasks require learners to use a variety of verb-based phraseological units, what is most salient to raters is how sophisticated the units are. In the LCF study, for example, the mean PMI of direct objects (and not diversity) was found to be significantly higher in C1 and C2 level essays and was an important predictor of proficiency in the random forest model. Likewise, in the TEF data, the PMI of verb-bundles (but not diversity) was a significant predictor of proficiency in both oral and written tasks. As argued in the discussion of that chapter, this may have been because both the letter to the editor and the telephone call task required the expression of stance and thus promoted the use of subordinate clauses (headed by verb-bundles). Interestingly, in the LANGSNAP data, it was also the PMI of direct objects that showed significant growth over time in the learner group, which possibly suggests that the ubiquity of direct objects made the use of sophisticated combinations more salient to learners and contributed to the incorporation of such units in their own production. The important

role of sophistication in relation to verb-based phraseological units in these studies lends further support to the argument of Biber et al. (2021) for more fine-grained complexity measures (i.e., those that take into account specific grammatical structures and functions) and shows, especially, how looking at their "co-occurrence with particular controlling verbs/nouns/adjectives" (Biber et al., 2020: 13) can perhaps be even more informative about proficiency and development than simplistic omnibus measures of (syntactic) complexity.

Putting everything together, the results of the three studies speak to the relevance of communicative function to phraseological complexity. Specifically, phraseological units that overlap with phrase-level syntactic complexity features were found to be more diverse in written tasks that required learners to incorporate information to form an argument. However, the opposite case was not found to be true: personal or involved tasks did not necessarily have more diversity in the use of verb-based phraseological units as such units tended to be common across all tasks. Rather, due perhaps to the fact that they were common across tasks, what seemed to be more salient to proficiency raters (and perhaps also learners themselves) was therefore not diversity of verb-based units but rather the sophistication of such units. In general, this is in line with previous register-based research across a variety of languages (Biber, 2014) but also shows how looking at complexity at the interface of traditional linguistic domains can be even more informative, or in other words, how "lexicogrammatical patterns [can] help us better understand the nature of complexity itself" (Biber et al., 2021: 431).

7.5 Limitations and implications for future research

It is important to keep in mind several limitations of this thesis when considering these results. First, on the level of methodology, as briefly discussed at the beginning of this chapter, the studies that comprise this thesis are based on different populations of language users. The LCF, LANGSNAP and TEF corpora include data from learners with dif-

ferent L1 backgrounds, different levels and types of exposure to French (including second- and well as foreign-language contexts) as well as different types of tasks. To a certain extent, the choice of these corpora was constrained by the availability of production data for learners of L2 French and especially a lack of corpora containing both oral and written data from the same learners, an essential aspect of this thesis. To these ends, it was necessary to use existing corpora despite the fact that the data may not have always been perfectly suited to the purposes of cross-modal comparison.

In the case of the LANGSNAP corpus, for example, although it contained oral and written data from the same learners, comparison of the three tasks was made difficult because they differ not only in modality but also communicative purpose (e.g., argumentative vs. narrative). This contrasts with the way that complexity has typically been compared across mode in previous research. As discussed in Section 2.3.2, researchers usually 'control' for task by having learners carry out the same task in both modalities. In one sense then, the fact that the LANGSNAP tasks differ in more than one way made it difficult to isolate the precise effects of modality compared to communicative purpose. By comparing the L1 and L2 data, the hope was that it would be possible to disentangle these two effects, with the assumption being that for the L1 group, processing would likely be somewhat more proceduralized than for the learner group so the effect of communicative purpose would be stronger. Whether this was a convincing argument is ultimately up to the reader. One could argue that it would have been better if the effects of task-type could be isolated so that any differences observed can more reliably be attributed to the effect of modality (cf. R. Ellis & Yuan, 2005). The fact is, however, that real-world tasks rarely differ only in terms of modality. If complexity measures are to be used as indices of proficiency or development, it is important to understand how both modality and communicative purpose interact. As shown throughout this thesis, the relationship between complexity, modality and development/proficiency is seldom straightforward. That being said, there is certainly space for replicating these results in more of an experimental setting in order to better control for task effects.

Another aspect that was difficult to control was topic. As shown by Paquot, Naets, et al. (2021), topic can have a strong effect on phraseological complexity. This was therefore a concern in each of the empirical studies presented in this thesis. In the case of the LCF study, the argumentative essays were written on one of seven different prompts. As shown in Table 7.7, while most topics exhibit a linear relationship between phraseological complexity and proficiency scores, there are differences between the topics in terms of phraseological complexity.

Table 7.7: Phraseological complexity measures across proficiency topics in LCF data

	Adjectival modifiers						Direct objects					
	RTTR			PMI			RTTR			PMI		
Topic	B2	C1	C2	B2	C1	C2	B2	C1	C2	B2	C1	C2
delinquency	4.96	4.71	5.39	2.21	2.23	3.02	4.64	4.85	5.24	1.77	1.85	1.78
euthanasia	2.34	2.82	3.36	1.88	3	2.71	3.53	3.43	3.62	1.72	1.98	2.35
foreigners	3.36	2.81	-	2.74	3.37	-	3.61	3.21	-	0.45	1.13	-
freedom	3.84	4	4.02	1.92	2.31	2.67	4.23	4.33	4.54	1.24	1.69	2.38
language	3.09	4.03	4.44	2.68	2.03	2.67	4.65	4.48	4.37	1.41	2.04	1.96
nuclear	3.13	2.74	-	4.72	5.45	-	1.89	2.91	-	1.59	2.26	-
state_reform	-	2.93	3.1	-	1.44	2.03	-	3.52	3	-	2.2	2.41

Note:
Dashed lines represent cases where there were no texts on a topic at a given proficiency level.
RTTR = Root type token ratio; PMI = Pointwise Mutual Information

For example, texts written on the topic of youth delinquency exhibited, on average, more adjectival modifier diversity than texts on the topic of euthanasia. That being said, because the LCF study only used one production per learner, topic differences were not overly problematic and could easily be controlled for by adding topic as a control variable in the random forest model.

In the LANGSNAP study, controlling for topic was more challenging because the topics of the essays as well as the pictures used in the oral narrative task were not consistent at each time point. This meant that differences between adjacent time points could have been due to topic differences rather than a learner's development. This required a more delicate solution than simply adding topic as a control variable

in the model.[9] An attempt was therefore made to control for topic by calculating the cosine similarity between the each text and that participant's first argumentative essay and using that measure as a control variable in the model. Of course, cosine similarity only provides a rather vague indication of the similarity of two texts because it is based solely on the overlap in individual words. The effect of topic can be much more subtle than differences in individual words however. For example, Paquot, Naets, et al. (2021) showed that certain prompts allowed learners to draw on larger semantic domains (see discussion in Section 2.3.3), which means despite the calculation of cosine similarity, there may still have been a confounding effect of topic in the LANGSNAP data.

The issue of topic was also a challenge in the TEF study where there were over one hundred different prompts for each task type. To account for this, the TEF study employed a topic modelling technique. In comparison to cosine similarity which is based on individual words, topic modelling works by looking at which words commonly group together across different texts so it provides a slightly more fine-grained quantification of topic than cosine similarity. Still, as a 'bag of words' technique, topic modelling remains quite limited because it has no way of capturing the relative differences in semantic domains that may be elicited by different topics. There is clearly scope for the exploration of more advanced techniques in order to control for topic within phraseological complexity research. This may be even more important for phraseological complexity than lexical or syntactic complexity given that the PMI identifies units that are used in specialized domains (a product of their exclusivity).

The focus on L2 French was also the source of some methodological difficulties. In addition to the scarcity of cross-modal learner corpora, it was also difficult to find an appropriate reference corpus for the calculation of PMI measures in these studies. There were three main criteria that were important in this regard. First, it was important to have a large enough corpus so that a sufficient number of phraseological units used by the learners would be attested. Second, for comparabil-

[9] Doing so would mean only being able to compare time points with the same prompt, masking development between adjacent time points.

ity with the results of Paquot (2018, 2019), it was important to find a reference corpus that was provided with dependency annotation[10] so that the same structures (direct objects, adjectival modifiers and adverbial modifiers) could be extracted (within the scope of the project, it was impossible to carry out this annotation step). Third, it was important that the reference corpus reflected the types of tasks in the learner corpora (see Larsson et al., 2021).

Unfortunately, it was not possible to find a reference corpus that matched all three of these criteria. To be sure, there are certainly large reference corpora for French, including the 9.8 billion word frTenTen2012 corpus (Jakubíček et al., 2013), the 1.5 billion word Leipzig corpus (Leipzig Corpora Collection, 2012) and the 1.2 billion word WaCky corpus (Baroni et al., 2009). However, unlike the FRCOW16 corpus (Schäfer, 2015; Schäfer & Bildhauer, 2012) that was ultimately used, these other French corpora do not contain dependency annotation.

The drawback of the FRCOW16 corpus is that, as a web-scraped corpus, it is not necessarily reflective of the types of tasks completed by the learners. Moreover, because of copyright protection, the corpus is distributed as randomized sentences, which makes it difficult to know the precise origin of each corpus document and thus the register it belongs to.

The fact that the FRCOW16 corpus only contains written data also means that it is not necessarily representative of oral French. Ideally, it would have been better to (additionally) use a reference corpus of oral French when comparing phraseological complexity across modes. Unfortunately, oral corpora of French are either not publicly available (in the case of New et al.'s 2007 corpus of French subtitles) or are much too small. The *Corpus de Français Parlé Parisien des années 2000* (CFPP, Branca-Rosoff et al., 2012), *Enquêtes sociolinguistiques à Orléans* (ESLO, Baude & Dugua, 2017) and *Corpus de langues parlées en intéraction* (CLAPI, Baldauf-Quilliatre et al., 2016), for example, only comprise 768- 28- and 16-thousand tokens respectively. Although the FRCOW16 corpus is not the best fit for the data in terms of register, its

[10] Even though the LANGSNAP and TEF studies did not use dependency labels, it was important to ensure comparability between these studies that the same reference corpus was used.

large size (10 billion words) means that the majority of the phraseological units used in the LCF and LANGSNAP corpora are attested (see Table 7.8). Unfortunately, even with a large reference corpus, there was a high percentage of unattested units in the TEF study. This is due in part to the fact that as units become longer, the chance of finding that exact unit in a reference corpus becomes lower.

In order to increase coverage as much as possible and to account for the highly inflected nature of French (Treffers-Daller, 2013) (an issue which is even more crucial when comparing across modes, given that many grammatical inflections are have a silent orthography in French), the units were lemmatized and proper nouns and determiners were collapsed together. However, these steps diverge somewhat from the typical practice in lexical bundle research which uses unlemmatized surface forms (Biber et al., 2004; Biber & Barbieri, 2007) and may have masked some important variation in lexical bundles. As an additional check, a search was run for the unattested units in the three oral corpora listed above. None of the unattested units were found in the CLAPI or ESLO corpus and only 26 unattested units were found in the CFPP corpus. In light of these results, a possible solution could have been to simply merge the CFPP corpus with the FRCOW16 corpus in order to better represent oral data. While this may appear at first glance to be an attractive solution, it would artificially raise the PMI of the units that only appear in the CFPP corpus relative to other units so ultimately it was decided to use only the FRCOW16 corpus, despite the fact that it may not be entirely representative. In the case of English, relatively large oral corpora are already available. For example, the 127-million spoken component of COCA means that oral-specific PMI measures can be calculated with the use of the TAALES tool (Kyle & Crossley, 2015). In order to get a clearer picture of oral phraseology in L2 French, there is a clear need for such a corpus in French.

Along the same lines, and as discussed in Chapter 3, there is also a need for more tools for the analysis of L2 French. This is perhaps most evident in the domain of syntactic complexity. While the development of the *fsca* tool did make it possible to calculate measures of syntactic complexity, the reliability of these measures was not always

Table 7.8: Percentage of units attested in FRCOW16 corpus

Unit	Types	Attested (%)	Unattested (%)
LCF			
Adverbial modifier	2770	2740 (98.92%)	30 (1.08%)
Adjectival modifier	3529	3376 (95.66%)	153 (4.34%)
Direct object	3936	3783 (96.11%)	153 (3.89%)
LANGSNAP			
Adjectival modifier	3979	3866 (97.16%)	113 (2.84%)
Direct object	5671	5558 (98.01%)	113 (1.99%)
TEF			
noun-bundle	78482	46595 (59.37%)	31887 (40.63%)
verb-bundle	71042	50343 (70.86%)	20699 (29.14%)

as high as would be desirable, especially compared to the high-levels of reliability reported for Lu's (2010) L2 Syntactic Complexity Analyzer (L2SCA). In addition, as discussed in Section 7.3.1, these omnibus measures collapse distinct syntactic structures together and are agnostic to the function that such structures play (cf. Biber et al., 2020, 2021). In the context of this thesis, the use of these measures made it possible to replicate Paquot (2018, 2019) and made the results somewhat comparable. In future work, it may be more informative to use syntactic complexity measures that target more specific structures (e.g., nominalizations, adverbial clauses, noun phrase adjuncts).

In addition, as shown by Biber et al. (2016), individual syntactic structures may not by themselves explain a large amount of variance in proficiency scores. Rather, it is the confluence of syntactic structures (e.g., those typical of literate discourse) that is more predictive of proficiency. For L2 English, these structures can be easily extracted using, for example Biber's (1991) *Variation across Speech and Writing* tagger, or the more recent *Multidimensional Analysis Tagger* (Nini, 2019). One tool that does exist for L2 French is *Direkt Profil* (Granfeldt et al., 2005; Granfeldt & Ågren, 2014), an automatic classifier which assigns texts to one of Bartning and Schlyter's (2004) developmental stages. However, this tool only focuses on the presence of morphosyntactic features within noun and verb phrases (e.g., determiner-noun agreement, percentage of different verb inflections) and therefore does not measure

syntactic complexity in the sense of Bulté and Housen's (2012) definition or that of Biber et al. (2020, 2021), which was why it was not used in this thesis.

Before more fine-grained tools can be developed for French however, it will be necessary to improve the reliability of basic NLP tools. As shown in Chapter 3, the output of Malt Parser and Treetagger was not always highly accurate on L2 French. The inaccuracy of these tools had an effect not just on the calculation of syntactic complexity measures but also on the extraction of phraseological units, given that identification of such units was dependent on lemmatization and POS-tagging. To this end, a useful avenue for future research would be to test the reliability and accuracy of more recently developed dependency parsing tools (e.g., spaCy, Honnibal & Montani, 2017) on L2 French. It is worth noting that colleagues at the UCLouvain are currently working on such a tool for readability assessment that may meet some of these needs (e.g., the calculation of more specific and accurate measures of syntactic complexity for French) (Wilkens et al., submitted).

Another limitation of this thesis was the focus on a restricted set of phraseological units in the LCF and LANGSNAP studies. The initial selection of these units was based on previous research showing that these structures are particularly difficult for L2 learners to master (e.g., Altenberg & Granger, 2001; Durrant & Schmitt, 2009; Laufer & Waldman, 2011; Nesselhauf, 2005). As discussed in the beginning of this chapter, one of the goals of the TEF study was to broaden the set of phraseological units under investigation by looking at four-word lexical bundles. This made it possible extract phraseological phenomena that were not captured in the LCF and LANGSNAP studies (e.g., *tenir_VER compte_NOM du_PRP point_NOM*). However, lexical bundles often extend beyond traditional syntactic boundaries (Biber et al., 1999). This meant that not all extracted units necessarily met Gries's (2008) criterion that phraseological units form "one semantic unit" (e.g., *rédacteur_NOM en_PRP chef_NOM je_PRO*, 'editor in chief I'). Along these lines, not all POS structures are necessarily equally informative about phraseological complexity and, by extension, proficiency/development. To illustrate, Table 7.9 lists the median PMI value for the ten most frequent POS structures identified in the TEF

Table 7.9: Median PMI across different POS-bundles

POS Structure	Types	Median PMI	IQR
NOM PRP DET NOM	5575	4.66	4.26
VER PRP DET NOM	5181	4.27	5.17
VER DET NOM PRP	4967	3.57	5.52
NOM PRO PRO VER	2644	3.07	3.91
VER VER DET NOM	2498	3.35	2.06
VER KON PRO VER	2330	4.18	3.96
VER PRP NOM PRP	2325	3.96	5.72
NOM PRP NOM PRP	2201	6.92	6.28
VER ADV PRP NOM	1798	4.78	1.80
VER DET NOM PRO	1746	2.75	4.31

corpus. What is evident is that there is quite some variation both between structures (shown by the differences in median PMI) and within structures (shown by the IQR). It is notable, in light of the claims of Vinay & Darbelnet (1995), that the two most frequent POS structures extracted from the TEF data are nouns and verbs modified by a prepositional phrase.

Taking a closer look at the units of this type with the highest PMI reveals that they tend to comprise idiomatic expressions (e.g., *boeuf_NOM avant_PRP DET charrue_NOM*, 'cart before the horse,' PMI = 18.63) which suggests that NOM/VER PRP DET NOM bundles may reveal more about a learner's phraseological sophistication than other types of four-word POS structures. A useful area of future research would be to explore this question in a more principled manner in order to determine which syntactic structures have phraseological patterns that are most informative about proficiency in L2 French.

This may appear, at first glance, to be a relatively straightforward proposition. In actuality however, it requires a subtle shift in the way that phraseological complexity research has been conducted up to this point. Until now, the focus has been on the extent to which measures of phraseological complexity, selected on the basis of their theoretical grounding, are predictive of proficiency and/or development. This, on its own cannot speak to the validity of such measures as indices of phraseological complexity however. As explained by Pal-

lotti (2015: 127), "if any reliable correlation were to be found between structural linguistic complexity and development, task conditions, or anything else, this should be considered more a validation of the theory postulating such a relationship rather than a validation of the complexity measure itself." In other words, the correlations found between phraseological complexity measures and proficiency or development in this thesis cannot speak to extent to which such measures really reflect the underlying construct of phraseological complexity. As argued by Paquot (2021), what is urgently needed in phraseological complexity research is to ground measures of phraseological complexity not just in human judgements of proficiency and development, but in judgements of the construct itself (e.g., asking raters to provide holistic impressions of a text's 'phraseological complexity'). This would be in line with similar work that has been done for the construct of lexical diversity (e.g., Jarvis, 2017). According to Paquot, having a better understanding of human perceptions of phraseological complexity would contribute to establishing the validity of phraseological complexity measures.

Finally, it almost goes without saying that in focusing solely on complexity, the current thesis offers only a limited view of the relationship between phraseology and L2 proficiency and development. This is clearly exemplified by the fact that although phraseological complexity measures were significant predictors of TEF exam scores, the fixed effects in the linear mixed effects model still only explained 33% of the variance. This should serve as a useful reminder that there are other factors beyond complexity that influence ratings of proficiency. In the case of L2 French, at least some of the variation in test scores is likely due to the presence of morphosyntactic deviances, which as discussed previously tend to persist at even very advanced levels of proficiency (Bartning et al., 2009; Forsberg & Bartning, 2010). This points to the importance of comparing the relative contributions of phraseological complexity and other dimensions of L2 performance (i.e., accuracy and fluency). The example of noun-bundle sophistication is illustrative in this respect. The PMI of noun-bundles was not found to be a significant predictor of proficiency in the TEF study. However, if it is indeed the case that learners make more morphosyntactic er-

rors within noun phrases than within verb phrases (as suggested by Forsberg & Bartning, 2010), then it is possible that a production with a relatively high level of phraseological complexity (but low accuracy) will receive a low score from raters. If 'accuracy' is controlled for in the analysis, phraseological complexity measures may therefore have more predictive strength. In a similar way, phraseological complexity is also likely to interact with fluency. As shown by early studies of phraseology in L2 French (e.g., Raupach, 1984; Towell et al., 1996), the use of certain 'formulaic' word combinations may increase a learner's rate of speech, which means that a production that has a relatively low level of phraseological complexity (because of the repetition of a small set of 'phraseological teddy bears') may receive a high score because it exhibits a high degree of fluency. While previous research has investigated phraseology in terms of other dimensions of L2 performance such as accuracy (Bulon & Meunier, 2020; Laufer & Waldman, 2011; Nesselhauf, 2005; Thewissen, 2008) or fluency (Boers et al., 2006; Osborne, 2007; Wood, 2009), attempts to measure the interaction of these dimensions is still rare. One exception is a study by Tavakoli & Uchihara (2020) who found that ngram association strength (defined in this thesis as a measure of phraseological *complexity*) was significantly associated with measures of oral fluency (higher frequency of end-clause pauses and lower frequency of repairs). The authors noted as well that lower-level learners used a more restricted set of ngrams compared to more advanced learners, which is generally in line with the results of the TEF study. These results suggest that it would be fruitful to investigate not only the direct relationship between phraseological complexity and L2 proficiency/development, but also how the relationship of phraseological complexity to proficiency or development may be mediated by accuracy or fluency.

7.6 Concluding remarks

As mentioned in the introduction, the original intent behind the use of complexity measures in L2 research was to find a "developmental yardstick against which global (i.e., not skill nor item specific) second language proficiency could be gauged" (Larsen-Freeman, 1983: 287).

Like a yardstick, the hope was to find an index of development and proficiency that would work across different contexts, including across different levels of proficiency or different target languages (Larsen-Freeman, 1976). After all, the usefulness of a yardstick depends, to a certain extent, on its *consistency* in different situations. The unit of length represented by an "inch" does not change depending on whether you are at the beginning or end of a construction project or whether you are building a simple birdhouse or a modern satellite. In fact, in some cases, a lack of measurement consistency can even be quite disastrous, as when a NASA contractor inadvertently used imperial rather than metric units, resulting in the loss of the Mars Climate Orbiter in 1999 (Stephenson et al., 1999). SLA research of course, is not rocket science (pardon the pun) and the interpretation of a complexity measure often depends very much on the context in which it is observed.

In the empirical studies presented in this thesis, contextual factors of both target language (French) and modality (oral versus written) were found to be highly relevant to understanding the relationship between complexity measures and proficiency/development. In Chapter 4, for example, it was found that both phraseological as well as non-phraseological complexity measures (i.e., lexical, syntactic, morphological) were important predictors of advanced proficiency, contributing to differences between B2, C1 and C2 level texts. This diverged somewhat with previous findings for L2 English, in which lexical or syntactic measures were less useful in distinguishing between the highest levels of proficiency than phraseological measures (Paquot, 2018, 2019).

The fact that French has a relatively richer verbal morphology system likely means that morphosyntactic aspects of a learner production continue to be important indicators of even highly advanced proficiency (Bartning et al., 2009; Forsberg & Bartning, 2010), in comparison to English where such factors may be less important at the highest levels. Another important finding was that the phraseological units that are most indicative of proficiency may also differ across languages. In French, for example, adverbial modifiers (e.g., *savoir + plus*, 'know + more') seem to be less informative about proficiency compared to ad-

jectival modifiers or direct objects. As argued by Vinay & Darbelnet (1995), French prefers verbal modification through the use of prepositional phrases rather than single word adverbs. Indeed, longer phraseological units including nouns and verbs modified by prepositional phrases (e.g., *tenir compte du point*) seemed to be particularly informative of proficiency in the TEF data. These two examples illustrate the importance of considering the linguistic characteristics of the target language when interpreting the meaning of a phraseological complexity measure.

Another contextual factor that was found to be important across the studies in this thesis was modality. Specifically, this was reflected in the way that oral and written tasks often differed in terms of both communicative function as well production circumstances. As shown in the LANGSNAP (Chapter 5) and TEF studies (Chapter 6), because noun-based phraseological units tend to be associated with phrasal elaboration, they are more frequent in tasks that promote the dense packaging of information (e.g., constructing a nuanced argument with many propositions), which may also be easier in the written mode because there is a relatively greater opportunity for online planning as compared to speech (Manchón, 2014; Skehan, 2014). As a result, noun-based phraseological complexity measures seem to be more informative about proficiency for written/informational type tasks than for more personal or involved oral tasks (e.g., narrating a picture story). This was exemplified in the TEF study, where it was found that the relationship between diversity of noun-based units and proficiency scores differed across modes: positive in the case of writing, negative for speech. Knowing the modality and communicative function of a task is thus key to being able to understand whether the phraseological diversity exhibited in production is indicative of higher or lower proficiency.

In addition to target language and modality, the interpretation of phraseological complexity measures was also found to be at least partially dependent on the proficiency level of the learners. This was shown by the fact that the learners in the LANGSNAP corpus exhibited about the same level of phraseological diversity as the L1 group in all tasks, which showed that they had likely already achieved

a ceiling level of phraseological complexity that was sufficient to complete the tasks so not much further development could really be expected. Only when a much broader range of proficiency levels was compared (in the case of the TEF study) was diversity found to be indicative of proficiency, in line with previous studies that have included lower level learners (below B2) (Garner, 2020; Jiang et al., 2021; Rubin et al., 2021).

As these examples illustrate, the relationship between phraseological complexity measures and proficiency/development appears to be very much context-dependent. This means that the interpretation of such measures needs to be driven by an understanding of the characteristics of the target language more broadly as well as the specific communicative task that a learner is performing. That such factors should to be taken into account, of course, is not a new idea. Norris and Ortega (2009: 556) argue for such an approach in CAF more generally:

> "[w]e would like to propose here that measurement practices in CAF must become considerably more organic, in the sense that they need to capture the fully integrated ecology of CAF development in specific learning contexts over time so as to help us understand how and why language develops or not within them."

The findings of this thesis point to the necessity of taking a more "organic" approach to phraseological complexity as well: one that recognizes the role of factors such as target language and modality as well as the fact that different aspects of complexity may be important at different stages of L2 development.

It is also important to recognize that compared to the long history of lexical or syntactic complexity measures in L2 research, the construct of phraseological complexity is still relatively new, and as such there are important questions that need to be answered. On a fundamental level, there is a need to test the validity of these measures according to human perceptions of phraseological complexity and not just proficiency (Paquot, 2021). As argued by Pallotti (2015, 2021), the association between complexity measures and proficiency scores or growth over

time is only informative about the usefulness of those measures as indices of proficiency or development, it does not reveal whether they indeed measure what they intend to measure. At the level of observation, it is important to determine the extent to which the constructs of 'diversity' and 'sophistication' are truly represented by the way they have thus far been operationalized (e.g., as RTTR, no. types, PMI). Likewise, in order to make inferences about the meaning of these constructs, there is a need to evaluate the extent to which they tap into aspects of system or structural complexity as opposed to relative complexity.

Another important question that remains to be answered is the extent to which phraseological complexity measures interact with other dimensions of L2 performance such as accuracy and fluency. Truly understanding the relationship between phraseological complexity and proficiency or development requires examining the "dynamic interrelationship of the different outcome measures" (Skehan, 1998: 179). After all, phraseological complexity does not arise in isolation but as the confluence of a learner's various linguistic subsystems coming together to complete a communicative task. As argued by Larsen-Freeman (2006: 592), "because language is complex, progress cannot be totally accounted for by performance in any one subsystem."

This thesis has begun to shed some light on one small part of this system, illuminating some aspects of the inter-relationship between phraseological complexity and modality on the one hand and proficiency and development on the other. In some ways, this has shown that relationships were less straightforward than expected but this should not necessarily be seen as a deficiency of the construct of phraseological complexity. After all, in the words of Rescher (1998: 20), "new sorts of facts always come to light and things get increasingly complicated. And those cognitive complexities clearly reveal…the complexities that characterize the subject matter at hand."

References

Abeillé, A., & Barrier, N. (2004). Enriching a French treebank. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 2233–2236.

Abeillé, A., Clément, L., & Toussenel, F. (2003). Building a treebank for French. In A. Abeillé (Ed.), *Treebanks. Text, speech and language technology* (pp. 165–187). Springer. https://doi.org/10.1007/978-94-010-0201-1_10

Ackermann, K., & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL) - A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235–247. https://doi.org/10.1016/j.jeap.2013.08.002

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, *30*(4), 535–556. https://doi.org/10.1177/0265532213489568

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, *67*(1), 180–208. https://doi.org/10.1111/lang.12232

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, *22*(2), 173–194. https://doi.org/10.1093/applin/22.2.173

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*(2), 145–167. https://doi.org/10.1075/sar.18001.arv

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*(2), 81–92. https://doi.org/10.1016/j.esp.2011.08.004

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). John Benjamins. https://doi.org/10.1075/lllt.32.05agr

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford University Press.

Baldauf-Quilliatre, H., Carvajal, I. C. de, Etienne, C., Jouin-Chardon, E., Teston-Bonnard, S., & Traverso, V. (2016). CLAPI, une base de données multimodale pour la parole en interaction : Apports et dilemmes. *Corpus*, *15*, 1–21. https://doi.org/10.4000/corpus.2991

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The waCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226. https://doi.org/10.1007/s10579-009-9081-4

Bartning, I., Arvidsson, K., & Forsberg Lundell, F. (2015). Complexity at the phrasal level in spoken L1 and very advanced L2 French. *Language, Interaction and Acquisition*, *6*(2), 181–201. https://doi.org/10.1075/lia.6.2.01bar

Bartning, I., Forsberg, F., & Hancock, V. (2009). Resources and obstacles in very advanced French: Formulaic language, information structure and morphosyntax. *EUROSLA Yearbook*, *9*(1), 185–211. https://doi.org/10.1075/eurosla.9.10bar

Bartning, I., & Schlyter, S. (2004). Itinéraires acquisitionnels et stades

de développement en français L2. *Journal of French Language Studies*, *14*(3), 281–299. https://doi.org/10.1017/S0959269504001802

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

Baude, O., & Dugua, C. (2017). Les ESLO, du portrait sonore au paysage digital. *Corpus- Bases, Corpus, Langage - UMR 7320*. https://halshs.archives-ouvertes.fr/halshs-01679544

Benevento, C., & Storch, N. (2011). Investigating writing development in secondary school learners of French. *Assessing Writing*, *16*(2), 97–110. https://doi.org/10.1016/j.asw.2011.02.001

Bernardini, P., & Granfeldt, J. (2019). On cross-linguistic variation and measures of linguistic complexity in learner texts : Italian, French and English. *International Journal of Applied Linguistics*, *29*(2), 211–232. https://doi.org/10.1111/ijal.12257

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, *69*, 65–78. https://doi.org/10.1016/j.system.2017.08.004

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28–41. https://doi.org/10.1016/j.jslw.2014.09.004

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. https://doi.org/10.1163/9789004361133_011

Biber, D. (1985). Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics*, *23*(2), 337–360. https://doi.org/10.1515/ling.1985.23.2.337

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, *62*(2), 384–414. https://doi.org/10.2307/414678

Biber, D. (1988). *Variation Across Speech and Writing* (First edition). Cambridge University Press.

Biber, D. (1991). *Variation Across Speech and Writing* (Paperback edition). Cambridge University Press.

Biber, D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, *15*(2), 133–163. https://doi.org/10.1080/01638539209544806

Biber, D. (2009). A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics*, *14*(3), 275–311. https://doi.org/10.1075/ijcl.14.3.08bib

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*, *14*(1), 7–34. https://doi.org/10.1075/lic.14.1.02bib

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*(3), 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D., & Conrad, S. (2019). *Register, Genre, and Style* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108686136

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

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. https://doi.org/10.1002/j.2333-8504.2013.tb02311.x

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*(1), 5–35. https://doi.org/10.5054/tq.2011.244483

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, *37*(5), 639–668. https://doi.org/10.1093/applin/amu059

Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes*, *46*, 1–15. https://doi.org/10.1016/j.jeap.2020.100869

Biber, D., Gray, B., Staples, S., & Egbert, J. (2021). *The register-functional approach to grammatical complexity: Theoretical foundation, descriptive research findings, application*. Routledge. https://doi.org/10.4324/9781003087991

Biber, D., Leech, G., Johansson, S., & Conrad, S. (1999). *Longman grammar of spoken and written English*. Longman.

Blanche-Benveniste, C. (1995). De la rareté de certains phénomènes syntaxiques en francais parlée. *French Language Studies*, *5*, 17–29.

Blanche-Benveniste, C., & Adam, J.-P. (1999). La conjugaison des verbes: Virtuelle, attestée, defective. *Recherches Sur Le Français Parlé*, *15*, 87–112.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a Lexical Approach to the test. *Language Teaching Research*, *10*(3), 245–261. https://doi.org/10.1191/1362168806lr195oa

Bolly, C. (2008). *Les unités phraséologiques : un phénomène linguistique complexe? Séquences (semi-) figées construites avec les verbes "prendre" et "donner" en français écrit L1 et L2 : Approche descriptive et acquisitionnelle*. [Doctoral Dissertation, UCLouvain]. https://dial.uclouvain.be/pr/boreal/object/boreal:19625/datastream/PDF_08/view

Bolly, C. (2009). The acquisition of phraseological units by advanced learners of French as an L2: High-frequency verbs and learner corpora. In E. Labeau & F. Myles (Eds.), *The advanced learner variety: The case of French* (pp. 199–220). Peter Lang.

Börner, W. (1987). Schreiben im Fremdsprachenunterricht: Überlegungen zu einem Modell. In W. Lörscher & R. Schulze (Eds.), *Perspectives on language in performance* (pp. 1336–1349). Gunter

Narr.

Branca-Rosoff, S., Fleury, S., Lefeuvre, F., & Pires, M. (2012). *Discours sur la ville: Présentation du Corpus de Français Parlé Parisien des années 2000 (CFPP2000).* http://cfpp2000.univ-paris3.fr/CFPP2000.pdf

Brezina, V., & Pallotti, G. (2019). Morphological complexity in written L2 texts. *Second Language Research*, *35*(1), 99–119. https://doi.org/10.1177/0267658316643125

Brown, J. D. (2014). Classical theory reliability. In A. J. Kunnen (Ed.), *The companion to language assessment* (pp. 1165–1181). Wiley-Blackwell.

Bulon, A., & Meunier, F. (2020). Comparing CLIL and non-CLIL learners' phrasicon in L2 Dutch: The (expected) winner does not take it all. *International Journal of Bilingual Education and Bilingualism*, 1–24. https://doi.org/10.1080/13670050.2020.1834502

Bulté, B. (2013). *The development of complexity in second language acquisition: A dynamic systems approach* [Doctoral Dissertation, Vrije Universiteit Brussel].

Bulté, B., & Housen, A. (2009). The development of lexical proficiency in L2 speaking and writing tasks by Dutch-speaking learners of French in Brussels [Conference presentation]. *Task Based Language Teaching Conference*.

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. https://doi.org/10.1075/lllt.32.02bul

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42–65. https://doi.org/10.1016/j.jslw.2014.09.005

Bulté, B., Housen, A., Pierrard, M., & Van Daele, S. (2008). Investigating lexical proficiency development over time - The case of Dutch-speaking learners of French in Brussels. *Journal of French Lan-*

guage Studies, *18*, 277–298. https://doi.org/10.1017/S0959269508003451

Candito, M., Crabbé, B., & Denis, P. (2010). Statistical French dependency parsing: treebank conversion and first results. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 1840–1847. https://hal.inria.fr/hal-00495196/

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

Carroll, J. B. (1964). *Language and Thought*. Prentice Hall.

Chambre de commerce et d'industrie de Paris. (2010). *TEF: Le Test d'évaluation de français de la Chambre et de l'industrie de Paris*.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1–25). Routledge. https://doi.org/10.4324/9780203937891

Chen, Y.-H., & Baker, P. (2010). Lexical Bundles in L1 and L2 Academic Writing. *Language Learning and Technology*, *14*(2), 30–49.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 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). John Benjamins. https://doi.org/10.1075/lllt.10.04cob

Cooper, T. C. (1976). Measuring written syntactic patterns of second language learners of German. *Journal of Educational Research*, *69*(5), 176–183. https://doi.org/10.1080/00220671.1976.10884868

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

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*.

Covington, M. A., & McFall, J. D. (2010). Cutting the gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. https://doi.org/10.1080/09296171003643098

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal (Complex Systems)*.

Dabrowska, E. (2014). Recycling utterances: A speaker's guide to sentence processing. *Cognitive Linguistics*, *25*(4), 617–653.

David, A. (2008). A developmental perspective on productive lexical knowledge in L2 oral interlanguage. *French Language Studies*, *18*, 315–331. https://doi.org/10.1017/S0959269508003475

De Bot, K. (1992). A bilingual production model: Levelt's 'speaking' model adapted. *Applied Linguistics*, *13*(1), 384–404. https://doi.org/10.4324/9781003060406-37

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. https://doi.org/10.1075/lllt.29.01deb

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*(1), 69–94. https://doi.org/10.1075/eurosla.15.03dec

De Clercq, B. (2016). *The development of linguistic complexity: A comparative study on L2 French and L2 English* [Doctoral Dissertation, Vrije Universiteit Brussel].

De Clercq, B., & Housen, A. (2017). A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and

diversity. *The 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. https://doi.org/10.1177/0267658316674506

De Cock, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair & M. Hundt (Eds.), *Corpus linguistics and linguistic theory* (pp. 51–68). Rodopi.

De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner english on computer* (pp. 67–79). Addison Wesley Longman. https://doi.org/10.4324/9781315841342-5

DeKeyser, R. (2007). Study abroad as foreign language practice. In R. DeKeyser (Ed.), *Practice in a second language* (pp. 208–226). Cambridge University Press. https://doi.org/10.1017/cbo9780511667275.012

Demeuse, M., Desroches, F., Casanova, D., Crendal, A., & Holle, A. (2010). Validation empirique d'un test de français langue étrangère en regard du Cadre européen commun de référence pour les langues. *22e Colloque International de l'association Pour Le développement Des méthodologies d'Évaluation En Éducation (ADMEE-Europe) - Evaluation Des Curriculums Et Des ProGrammes d'éducation Et de Formation*, 881–902.

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). Brill. 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*, 721–736. https://doi.org/10.1007/s10579-012-9193-0

Duneton, C. (1973). *Parler croquant*. Stock.

Dunn, J. (2018). Multi-unit association measures: Moving beyond pairs

of words. *International Journal of Corpus Linguistics*, *23*(2), 183–215. https://doi.org/10.1075/ijcl.16098.dun

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*(2), 157–177. https://doi.org/10.1515/iral.2009.007

Edmonds, A., & Gudmestad, A. (2021). Collocational development during a stay abroad. *Languages*, *6*(1), 1–17. https://doi.org/10.3390/languages6010012

Eguchi, M., & Kyle, K. (2020). Continuing to explore the multidimensional nature of lexical sophistication : The case of oral proficiency interviews. *The Modern Language Journal*, *104*(2), 381–400. https://doi.org/10.1111/modl.12637

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*(1), 375–396. https://doi.org/10.1515/CLLT.2009.003

Ellis, N. C., & Wulff, S. (2015). Usage-Based Approaches to SLA. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction*. Routledge.

Ellis, N. C., & Wulff, S. (2019). Cognitive Approaches to Second Language Acquisition. In J. W. Schwieter & A. G. Benati (Eds.), *The Cambridge handbook of language learning* (pp. 41–61). Cambridge University Press. https://doi.org/10.1017/9781108333603.003

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford University Press.

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. https://doi.org/https://doi.org/10.1017/9781108643689.007

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. https://doi.org/10.1075/lllt.11.11ell

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

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, *20*(1), 29–62. https://doi.org/10.1515/text.1.2000.20.1.29

Ferrari, S., & Nuzzo, E. (2009). Meeting the challenge of diversity with TBLT: Connecting speaking and writing in mainstream classrooms [Conference presentation]. *Task Based Language Teaching Conference*.

Flower, L., & Hayes, J. R. (1980). The cognition of discovery: Defining a rhetorical problem. *College Composition and Communication*, *31*(1), 21. https://doi.org/10.2307/356630

Forsberg, F. (2008). *Le langage préfabriqé: Formes fonctions et fréquences en français parlé L2 et L1*. Peter Lang.

Forsberg, F. (2010). Using conventional sequences in L2 French. *IRAL - International Review of Applied Linguistics in Language Teaching*, *48*(1), 25–51. https://doi.org/10.1515/iral.2010.002

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, F., & Fant, L. (2010). Idiomatically speaking: Effects of task variation on formulaic language in highly proficient users of L2 French and Spanish. In D. Wood (Ed.), *Perspectives on formulaic language : Acquisition and communication* (pp. 47–40). Continuum.

Forsberg Lundell, F., Bartning, I., Engel, H., Gudmundson, A., Hancock, V., & Lindqvist, C. (2014). Beyond advanced stages in high-level spoken L2 French. *Journal of French Language Studies*, *24*(2), 255–280. https://doi.org/10.1017/S0959269513000057

Forsberg Lundell, F., & Lindqvist, C. (2014). Vocabulary aspects of advanced L2 French: Do lexical formulaic sequences and lexical richness develop at the same rate? In C. Lindqvist & C. Bardel (Eds.), *The acquisition of French as a second language: New developmental perspectives* (pp. 75–94). John Benjamins. https://doi.org/10.1075/bct.62.05for

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. *The Canadian Modern Language Review*, *74*(4), 627–649. https://doi.org/10.3138/cmlr.2017-0093

Fry, D. B. (1969). The linguistic evidence of speech errors. *BNRO Studies of English*, *8*, 69–74. https://doi.org/10.1515/9783110888423.157

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. *Language Learning*, *67*(1), 155–179. https://doi.org/10.1111/lang.12225

Gablasova, D., Brezina, V., & McEnery, T. (2019). The Trinity Lancaster Corpus Development, description and application. *International Journal of Learner Corpus Research*, *5*(2), 126–158. https://doi.org/10.1075/ijlcr.19001.gab

Galbraith, D. (2009). Writing as discovery. *British Journal of Educational Psychology*, *2*(6), 5–26.

Garner, J. (2020). The cross-sectional development of verb-noun collocations as constructions in L2 writing. *IRAL - International Review of Applied Linguistics in Language Teaching*, 1–27. https://doi.org/10.1515/iral-2019-0169

Garner, J., & Crossley, S. A. (2018). A latent curve model approach to studying L2 n-gram development. *The Modern Language Journal*,

102(3), 494–511. https://doi.org/10.1111/modl.12494

Garner, J., Crossley, S. A., & Kyle, K. (2018). Beginning and intermediate L2 writer's use of N-grams: An association measures study. *IRAL - International Review of Applied Linguistics in Language Teaching*, *58*(1). https://doi.org/10.1515/iral-2017-0089

Garner, J., Crossley, S. A., & Kyle, K. (2019). N-gram measures and L2 writing proficiency. *System*, *80*, 176–187. https://doi.org/10.1016/j.system.2018.12.001

Garretson, G. (2011). *Dexter Coder (Version 0.6.4)*. http://www.dextercoder.org/

Garrett, M. F. (1976). Syntactic processes in sentence production. In G. Bower (Ed.), *Psychology of learning and motivation* (Vol. 9, pp. 231–256). Academic Press.

Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 oral production. *IRAL - International Review of Applied Linguistics in Language Teaching*, *45*(3), 215–240. https://doi.org/10.1515/iral.2007.010

Goldberg, A. E. (2006). *Constructions at work*. Oxford University Press.

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.

Granfeldt, J., & Ågren, M. (2014). SLA developmental stages and teachers' assessment of written French: Exploring Direkt Profil as a diagnostic assessment tool. *Language Testing*, *31*(3), 285–305.

Granfeldt, J., Nugues, P., Persson, E., Persson, L., Kostadinov, F., Ågren, M., & Schlyter, S. (2005). Direkt Profil: A system for evaluating texts of second language learners of French based on developmental sequences. *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, 53–60.

Granger, S. (1998). *Learner English on Computer*. Longman.

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

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 27–50). John Benjamins. https://doi.org/10.1075/z.139.07gra

Green, N. (2011). Dependency Parsing. *Proceedings of the 20th Annual Conference of Doctoral Students*, 137–142.

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*. https://cran.r-project.org/package=vip

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. https://doi.org/10.1075/z.139.06gri

Gries, S. Th., & Ellis, N. C. (2015). Statistical measures for Usage-Based Linguistics. *Language Learning*, *65*(1), 228–255. https://doi.org/10.1111/lang.12119

Guiraud, P. (1954). *Les charactères statistiques du vocabulaire*. 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), 1–30. https://doi.org/10.1075/eurosla.14.01gyl

Hakuta, K. (1975). Learning to speak a second language: What exactly does a child learn? In P. Dato, Daniel (Ed.), *Georgetown university round table on languages and linguistics* (pp. 193–207). Georgetown University Press.

Halliday, M. A. K. (1966). Lexis as a linguistic level. In C. E. Bazell, J. C.

Catford, J. R. Firth, M. A. K. Halliday, & R. H. Robins (Eds.), *In memory of J.R. Firth* (pp. 148–=162). Longmans.

Halliday, M. A. K. (1989). *Spoken and written language*. Oxford University Press.

Halliday, M. A. K. (2004). *The language of science - the collected works of M.A.K. Halliday*. Continuum.

Halliday, M. A. K., & Martin, J. R. (1993). *Writing science*. The Falmer Press.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (1999). *Construing experience through meaning: A language-based approach to cognition*. Cassell.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2006). *Construing experience through meaning: A language-based approach to cognition*. Continuum.

Harley, B., & King, M. L. (1989). Verb lexis in the written compositions of young L2 learners. *Studies in Second Language Acquisition*, *11*(4), 415–439. https://doi.org/10.1017/S0272263100008421

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: a study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, *4*(2), 237–258. https://doi.org/10.1111/j.1473-4192.1994.tb00065.x

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, *29*(3), 369–388. https://doi.org/10.1177/0741088312451260

Hebb, D. O. (1949). *The organization of behavior*. Wiley.

Holliday, A. (2006). Native-speakerism. *ELT Journal*, *60*(4), 385–387. ht tps://doi.org/10.1093/elt/ccl030

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. (2006). Survival Ensembles. *Biostatistics*, *7*(3), 355–373. https: //doi.org/10.1093/biostatistics/kxj011

Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. (2022). *Documentation for the 'party' package*. https://cran.r-project.org/web/packages/party/party.pdf

Housen, A., De Clercq, B., Kuiken, F., & Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second Language Research*, *35*(1), 3–21. https://doi.org/10.1177/0267658318809765

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461–473. https://doi.org/10.1093/applin/amp048

Housen, A., Kuiken, F., & Vedder, I. (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. John Benjamins. https://doi.org/10.1075/lllt.32

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, *19*(1), 24–44. https://doi.org/10.1093/applin/19.1.24

Hulstijn, J. H., & de Graaff, R. (1994). Under what conditions does explicit knowledge of a second language facilitate the acquisition of implicit knowledge?: A research proposal. *AILA Review*, *11*, 97–112.

Hunston, S., & Francis, G. (2000). *Pattern Grammar*. John Benjamins.

Hunt, K. (1965). *Grammatical structures written at three grade levels*. NCTE.

Hunt, K. (1970). Do sentences in the second language grow like those in the first? *TESOL Quarterly*, *4*(3), 195–202.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen Corpus Family. *Proceedings from the 7th International Corpus Linguistics Conference*, 125–127.

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, *34*(4), 537–553. https://doi.org/10.1177/0265532217710632

Jiang, J., Bi, P., Xie, N., & Liu, H. (2021). Phraseological complexity and low- and intermediate-level L2 learners' writing quality. *International Review of Applied Linguistics in Language Teaching*. https:

//doi.org/10.1515/iral-2019-0147

Johnson, W. (1944). Studies in Language Behavior: I. A Program of Research. *Psychological Monographs*, *56*, 1–15.

Kane, M. (2006). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131–153). Lawrence Erlbaum. https://doi.org/10.4324/9780203874776.ch7

Kaszubski, P. (2000). *Selected aspects of the lexicon, phraseology and style in the writing of polish advanced learners of English: A contrastive, corpus-based approach* [Doctoral dissertation, Adam Mickiewicz University].

Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences and applications* (pp. 57–71). Routledge.

Kempen, G., & Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive Science*, *11*(2), 201–258. https://doi.org/10.1016/S0364-0213(87)80006-X

Kern, R. G., & Schultz, J. M. (1992). The effects of composition instruction on intermediate level French students' writing performance: Some preliminary findings. *The Modern Language Journal*, *76*(1), 1–13. https://doi.org/10.1111/j.1540-4781.1992.tb02572.x

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*(1), 120–141. https://doi.org/10.1111/modl.12447

Klein, D., & Manning, C. (2003). Fast exact inference with a factored model for natural language parsing. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 3–10). MIT Press.

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*(1), 60–69. https://doi.org/10.7820/vli.v01.1.koizumi

Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum. https://doi.org/10.4324/9780203763964

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.

Kormos, J., & Trebits, A. (2012). The role of task complexity, modality, and aptitude in narrative task performance. *Language Learning*, *62*(2), 439–472. https://doi.org/10.1111/j.1467-9922.2012.00695.x

Kremmel, B., Brunfaut, T., & Alderson, J. C. (2017). Exploring the role of phraseological knowledge in foreign language reading. *Applied Linguistics*, *38*(6), 848–870. https://doi.org/10.1093/applin/amv070

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, *17*(1), 48–60. https://doi.org/10.1016/j.jslw.2007.08.003

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

Kuiken, F., & Vedder, I. (2012). Syntactic complexity, lexical variation and accuracy as a function of task complexity and proficiency level in L2 writing and speaking. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 143–170). John Benjamins. https://doi.org/10.1075/lllt.32.07kui

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., & Crossley, S. A. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*(4), 513–535. https://doi.org/10.1177/0265532217712554

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, *102*(2), 333–349. https://doi.org/10.1111/modl.12468

Labbé, D. (2007). Coordination et subordination en français oral. *Les Nouvelles Journées de l'ERLA*, *4*, 161–182.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Larsen-Freeman, D. (1976). Evidence of the need for a second language acquisition index of development. In W. C. Ritchie (Ed.), *Principles of second language learning and teaching*. Academic Press.

Larsen-Freeman, D. (1978). An ESL Index of Development. *TESOL Quarterly*, *12*(4), 439–448.

Larsen-Freeman, D. (1983). Assessing global second language proficiency. In H. W. Seigler & M. H. Long (Eds.), *Classroom-oriented research in second language acquisition* (pp. 287–304). Newbury House.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, *27*(4), 590–619. https://doi.org/10.1093/applin/aml029

Larsen-Freeman, D., & Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning*, *27*(1), 123–134. https://doi.org/10.1111/j.1467-1770.1977.tb00296.x

Larsson, T., Paquot, M., & Biber, D. (2021). On the importance of register in learner writing: A multi-dimensional approach. In E. Seoane & D. Biber (Eds.), *Corpus-based approaches to register variation* (pp. 235–258). John Benjamins. https://doi.org/10.1075/scl.103.09lar

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*(2), 647–672. https://doi.org/10.1111/j.1467-9922.2010.00

621.x

Leech, G. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning*, *50*(4), 675–724. https://doi.org/10.1111/0023-8333.00143

Leipzig Corpora Collection. (2012). *French mixed corpus based on material from 2012*. https://corpora.uni-leipzig.de?corpusId=fra_mixed_2012

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. The MIT Press.

Levelt, W. J. M. (1999). Producing spoken language: A blueprint of the speaker. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 83–122). Oxford University Press. https://doi.org/10.1093/acprof

Levshina, N. (2015). Conditional inference trees and random forests. In *How to do linguistics with R: Data exploration and statistical analysis* (pp. 291–300). John Benjamins. https://doi.org/10.1075/z.195

Lindqvist, C., Bardel, C., & Gudmundson, A. (2011). Lexical richness in the advanced learner's oral production of French and Italian L2. *IRAL - International Review of Applied Linguistics in Language Teaching*, *49*(3), 221–240. https://doi.org/10.1515/iral.2011.013

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)* (pp. 109–126). European Second Language Association.

Lissón, P., & Ballier, N. (2018). Investigating lexical progression through lexical diversity metrics in a corpus of French L3. *Discours*, *23*. https://doi.org/10.4000/discours.9950

Lonsdale, D., & Le Bras, Y. (2009). *A frequency dictionary of French: Core vocabulary for learners*. Routledge.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36–62. https://doi.org/10.5054/tq.2011.240859

Macaro, E., & Masterman, L. (2006). Does intensive explicit grammar instruction make all the difference? *Language Teaching Research*, *10*(3), 297–327. https://doi.org/10.1191/1362168806lr197oa

MacWhinney, B. (2020). *The CHILDES Project Tools for Analyzing Talk–Electronic Edition Part 1: The CHAT Transcription Format*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.359.4451

Manchón, R. M. (2014). The internal dimension of tasks: The interaction between task factors and learner factors in bringing about learning through writing. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 27–52). John Benjamins.

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488. https://doi.org/10.1177/0265532207080767

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*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McManus, K., Mitchell, R., & Tracy-Ventura, N. (2021). A longitudinal study of advanced learners' linguistic development before, during, and after study abroad. *Applied Linguistics*, *42*(1), 1–29. https://doi.org/10.1093/applin/amaa003

Michalke, M. (2019). *koRpus: An R Package for Text Analysis (Version 0.12-1)*.

Michel, M., Murakami, A., Alexopoulou, T., & Meurers, D. (2019). Effects of task type on morphosyntactic complexity across proficiency: Ev-

idence from a large learner corpus of A1 to C2 writings. *Instructed Second Language Acquisition*, *3*(2), 124–152. https://doi.org/10.1558/isla.38248

Michel, M., Révész, A., Lu, X., Kourtali, N. E., Lee, M., & Borges, L. (2020). Investigating L2 writing processes across independent and integrated tasks: A mixed-methods study. *Second Language Research*, *36*(3), 307–334. https://doi.org/10.1177/0267658320915501

Miestamo, M. (2008). Grammatical complexity in cross-linguistic perspective. In M. Miestamo, K. Sinnemäki, & F. Karlson (Eds.), *Language complexity* (pp. 23–41). John Benjamins. https://doi.org/10.1075/slcs.94.04mie

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*(2), 243–277. https://doi.org/10.3366/cor.2017.0118

Myles, F., & Mitchell, R. (2011). *French Learner Language Oral Corpora*. http://www.flloc.soton.ac.uk/

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. John Benjamins.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*, 661–677. https://doi.org/10.1017/S014271640707035X

Nihalani, N. K. (1981). The quest for the L2 index of development. *RELC Journal*, *12*(2), 50–56.

Nini, A. (2019). The Multi-Dimensional Analysis Tagger. In T. Berber Sardinha & M. Veirrano Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 67–94). Bloomsbury. https://doi.org/10.5040/9781350023857.0012

Nivre, J., Hall, J., & Nilsson, J. (2006). MaltParser : A data-driven parser-generator for dependency parsing. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2216–2219.

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. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college level L2 writing. *Applied Linguistics*, *24*(4), 492–518. https://doi.org/10.1093/applin/24.4.492

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). de Gruyter.

Ortega, L. (2019). SLA and the study of equitable multilingualism. *Modern Language Journal*, *103*(1), 23–38. https://doi.org/10.1111/modl.12525

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. https://doi.org/10.1017/S0267190505000024

Osborne, J. (2007). Investigating L2 fluency through oral learner corpora. In M. C. Campoy & M. J. Luzón (Eds.), *Spoken corpora in applied linguistics* (pp. 181–198). Peter Lang.

Ovtcharov, V., Cobb, T., & Halter, R. (2006). La richesse lexicale des productions orales: Mesure fiable du niveau de compétence langagière. *The Canadian Modern Language Review*, *63*(1), 107–125. https://doi.org/10.3138/cmlr.63.1.107

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601. https://doi.org/10.1093/applin/amp045

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*(11), 117–134. https://doi.org/10.1177/0267658314536435

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*(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. (2021). Measures of phraseological complexity: Reliability and validity [conference presentation]. *World Congress of Applied Linguistics*.

Paquot, M., Gablasova, D., Brezina, V., & Naets, H. (in press). 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*. John Benjamins.

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

Paquot, M., Gries, S. Th., & Yoder, M. (2021). Measuring Lexicogrammar. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 223–232). Routledge. https://doi.org/10.4324/9781351034784-25

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.

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

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2020). *nlme: Linear and Nonlinear Mixed Effects Models*. 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*(2), 538–553. https://doi.org/10.1111/modl.12335

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 University Press.

Poulisse, N., & Bongaerts, T. (1994). First language use in second language production. *Applied Linguistics*, *15*(1), 36–57. https://doi.org/10.1093/applin/15.1.36

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research. *Language Learning*, *65*(1), 37–75. https://doi.org/10.1111/lang.12112

Qi, Y., & Ding, Y. (2011). Use of formulaic sequences in monologues of Chinese EFL learners. *System*, *39*(2), 164–174. https://doi.org/10.1016/j.system.2011.02.003

R Core Team. (2019). *R: A language and environment for statistical computing (Version 3.6.1)*. R Foundation for Statistical Computing. https://www.r-project.org/

R Core Team. (2021). *R: A language and environment for statistical computing (Version 4.1.0)*. R Foundation for Statistical Computing. https://www.r-project.org/

Raupach, M. (1984). Formulae in second language speech production. In H. W. Dechert, D. Mohle, & M. Raupach (Eds.), *Second language productions* (pp. 114–137). Gunter Nar.

Ravid, D., & Tolchinsky, L. (2002). Developing linguistic literacy: a comprehensive model. *Journal of Child Language*, *29*(2), 417–447. https://doi.org/10.1017/s0305000902005111

Read, J. (2000). *Assessing Vocabulary*. Cambridge University Press. https://doi.org/10.1017/cbo9780511732942

Rescher, N. (1998). *Complexity: A Philosophical Overview*. Transaction Publishers.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning*, *45*(1), 99–140. https://doi.org/10.1111/j.1467-1770.1995.tb00964.x

Robinson, P. (2001). Task complexity, cognitive resources and syllabus design: a triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287–318). Cambridge University Press.

Robinson, P. (2011). Second language task complexity, the Cognition Hypothesis, language learning, and performance. In *L2 task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 3–37). https://doi.org/10.1515/iral.2007.007

Robinson, P. (2015). The Cognition Hypothesis, second language task demands, and the SSARC model of pedagogic task sequencing. In M. Bygate (Ed.), *Domains and directions in the development of TBLT: A decade of plenaries from the international conference* (pp. 87–122). John Benjamins. https://doi.org/10.1075/tblt.8.04rob

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). Russell Sage Foundation.

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, *7*, 140–162. https://doi.org/10.1075/arcl.7.06rom

Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking. *Language Testing*, *34*(4), 477–492. https://doi.org/10.1177/0265532217711431

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. https://doi.org/10.21832/9781788924863-006

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 the third workshop on challenges in the management of large corpora (CMLC-3)* (pp. 28–34). Institute fur 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 (ELRA).

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *The Public Opinion Quarterly*, *19*(3), 321–325.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, *7*(3), 301–317.

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

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review*, *64*(3), 429–458. https://doi.org/10.3138/cmlr.64.3.429

Siyanova-Chanturia, A. (2015). Collocation in beginner learner writing: A longitudinal study. *System*, *53*, 148–160. https://doi.org/10.1016/j.system.2015.07.003

Siyanova-Chanturia, A., & Spina, S. (2020). Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study. *Language Learning*, *70*(2), 420–463. https://doi.org/10.1111/lang.12383

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Second Language Task-Based Performance*, *17*(1), 13–34.

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

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*, 1–14.

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. https://doi.org/10.1057/9780230242258

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

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. https://doi.org/10.1075/tblt.5.08ske

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.

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

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, *12*(3), 214–225. https://doi.org/10.1016/j.jeap.2013.05.002

Stefanowitsch, A., & Gries, S. Th. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *8*(2), 209–243. https://doi.org/10.1075/ijcl.8.2.03ste

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*(4), 321–343. https://doi.org/10.1515/iral.2011.017

Stephenson, A. G., Mulville, D. R., Bauer, F. H., Dukeman, G. A., Norvig, P., LaPiana, L. S., Sackheim, R., Rutlege, P. J., & Folta, D. (1999). *Mars Climate Orbiter Mishap Investigation Board Phase I Report (NASA)*.

Strobl, C., Boulesteix, A.-L., Kneib, T., Zeileis, T., & Achim, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(307), 1–11. https://doi.org/10.1186/1471-2105-9-307

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), 1–21. https://doi.org/10.1186/1471-2105-8-25

Szmrecsanyi, B., & Kortman, B. (2009). Between simplification and complexification: non-standard varieties of English around the world. In G. Sampson, D. Gil, & P. Trudgill (Eds.), *Language complexity as an evolving variable* (pp. 64–79). Oxford University Press.

Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning*, *70*(2), 506–547. https://doi.org/10.1111/lang.12384

Templin, M. (1957). *Certain language skills in children: Their development and interrelationships*. The University of Minnesota Press.

Thewissen, J. (2008). The phraseological errors of French-, German- and Spanish-speaking EFL learners: evidence from an error-tagged learner corpus. In A. F. Garcia, T. Rkibi, M. R. Cruz, R. Carvalho, C. Direito, & D. D. Santos (Eds.), *Proceedings of the*

eighth teaching and language corpora conference (TALC 08) (pp. 300–306). Associação de Estudos e de Investigação Científica do ISLA-Lisboa.

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

Tidball, F., & Treffers-Daller, J. (2007). Exploring measures of vocabulary richness in semi-spontaneous French speech. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 133–149). Cambridge University Press.

Tidball, F., & Treffers-Daller, J. (2008). Analysing lexical richness in French learner language: What frequency lists and teacher judgements can tell us about basic and advanced words. *Journal of French Language Studies*, *18*(3), 299–313. https://doi.org/10.1017/S0959269508003463

Towell, R. (2012). Complexity, accuracy and fluency from the perspective of psycholinguistic second language acquisition research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency : Complexity, accuracy and fluency in SLA* (pp. 47–70). John Benjamins. https://doi.org/10.1075/lllt.32.03tow

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, *17*(1), 84–119. https://doi.org/10.1093/applin/17.1.84

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. https://doi.org/10.21832/9781783092291-011

Tracy-Ventura, N., Mitchell, R., & McManus, K. (2016). The LANGSNAP longitudinal learner corpus: Design and use. In U. Römer & E. Tognini-Bonelli (Eds.), *Spanish learner corpus research : Current trends and future perspectives* (pp. 117–142). John Benjamins. https://doi.org/10.1075/scl.78.05tra

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.

Tutin, A., & Grossman, F. (2014). *L'écrit scientifique : du lexique au discours. Autour de Scientext*. Presses Universitaires de Rennes.

Uchihara, T., Eguchi, M., Clenton, J., & Saito, K. (2021). To what extent is collocation knowledge associated with oral proficiency ? A Corpus-based approach to word association. *Language and Speech*, *65*(2). https://doi.org/10.1177/00238309211013865

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

Vanderbauwhede, G. (2012). *Le déterminant démonstratif en français et en néerlandais à travers les corpus : Théorie, description, acquisition* [Doctoral Dissertation, KULeuven and Université Paris Ouest Nanterre La Défense].

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*(2), 197–229. https://doi.org/10.1075/ijlcr.20015.van

Vandeweerd, N., Housen, A., & Paquot, M. (in revisions). Comparing the longitudinal development of phraseological complexity across oral and written tasks. *Studies in Second Language Acquisition*.

Vasylets, O., Gilabert, R., & Manchón, R. M. (2017). The effects of mode and task complexity on second language production. *Language Learning*, *67*(7), 394–430. https://doi.org/10.1111/lang.12228

Vasylets, O., Gilabert, R., & Manchón, R. M. (2019). Differential contribution of oral and written modes to lexical, syntactic and propositional complexity in L2 performance in instructed contexts. *Instructed Second Language Acquisition2*, *3*(2), 206–227. https://doi.org/10.1558/isla.38289

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.). John Benjamins.

Wang, J., & Dong, Y. (2020). Measurement of text similarity: A survey. *Information*, *11*(9), 1–17. https://doi.org/10.3390/info11090421

Way, D. P., Joiner, E. G., & Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *The Modern Language Journal*, *84*(2), 171–184. https://doi.org/10.1111/0026-7902.00060

Welcomme, A. (2013). Jonction interpropositionnelle et complexité syntaxique dans les récits d'apprenants néerlandophones et locuteurs natifs du français. In U. Paprocka-Piotrowska, C. Martinot, & S. Gerolimich (Eds.), *La complexité en langue et son acquisition* (pp. 261–284). Towarzystwo Naukowe Kul Katolicki Uniwersytet Jana Pawla II.

Wilkens, R., Alfter, D., Wang, X., Pintard, P., Tack, A., Yancey, K., & François, T. (submitted). FABRA: French Aggregator-Based Readability Assessment toolkit. In *Proceedings of the thirteenth international conference on language resources and evaluation (LREC'22)*.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy & complexity*. Second Language Teaching & Curriculum Center.

Wood, D. (2009). Effects of Formulaic Sequences on Fluency: A case study. *Canadian Journal of Applied Linguistics*, *12*, 39–57.

Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics*, *21*(4), 463–489. https://doi.org/10.1093/applin/21.4.463

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. https://doi.org/10.1016/j.jslw.2016.11.001

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*(2), 236–259. https://doi.org/10.1093/applin/amp024

Zalbidea, J. (2017). 'One Task Fits All'? The roles of task complexity, modality, and working memory capacity in L2 performance. *The Modern Language Journal*, *101*(2), 335–352. https://doi.org/10.1111/modl.12389

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. https://doi.org/10.1016/j.system.2020.102437

Zhang, X., Zhao, B., & Li, W. (2021). N-gram use in EFL learners' retelling and monologic tasks. *IRAL - International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2021-0080

Appendix

Appendix I

Definitions of syntactic units used in fsca tool

Sentences

Following Lu (2010: 481) we defined a sentence as "a group of words delimited by one of the following punctuation marks that signal the end of a sentence: period, question mark, exclamation mark, quotation mark or ellipsis." Because the CONLL input to the `getUnits()` function is already segmented into sentences, no additional query is required.

Clauses

Clauses are defined as structures with a subject and a finite verb (Hunt, 1965). After identifying dependent clauses and T-units, clauses include all T-units as well as all subordinate clauses embedded within the T-units in a given sentence.

Dependent clauses

Dependent clauses are clauses which are semantically and/or structurally dependent on a super-ordinate syntactic structure. They include nominal clauses, adverbial clauses and adjectival clauses (Hunt, 1965; Lu, 2010). They must contain a finite verb and a subject.

Nominal clauses are extracted using subordinate conjunctions (e.g., *que*, 'which') as the main node.

Ils prétendent qu'il est impossible de rééduquer un tel jeune criminel.

```
getUnits(test.sents[["b.208.1"]],
         what = "tokens",
         units = c("DEP_CLAUSES"),
         paste.tokens = TRUE)
```

- qu' il est impossible de rééduquer un tel jeune criminel

Adverbial clauses are also extracted using subordinate conjunctions as the main node.

Quand les lois sont contre le droit, il n'y a qu'une héroïque façon de protester contre elles : les violer (Hugo).

```
getUnits(test.sents[["b.347.1"]],
         what = "tokens",
         units = c("DEP_CLAUSES"),
         paste.tokens = TRUE)
```

- Quand les lois sont contre le droit

Adjectival clauses are extracted using pronouns and finite verbs that modify a noun as the main nodes (e.g., *durent* and *arrivent* in the example below).

Des procès qui durent des années, des déclarations d'impôts inhumainement longues et compliquées, un gouvernement qui n'arrive pas à prendre forme...

```
getUnits(test.sents[["b.110.1"]],
         what = "tokens",
         units = c("DEP_CLAUSES"),
         paste.tokens = TRUE)
```

- qui durent des années
- qui n' arrive pas à prendre forme

Special cases

Clauses of the type 'il y a' are considered dependent clauses only if the verb *avoir* is directly dependent on a finite verb or if there is a finite verb dependent on *avoir*. This means that adverbial clauses which function like 'ago' in English (e.g., *il y a deux ans...*; 'two years ago...') are considered dependent clauses but simple declaratives (e.g., *il y a une maison*; 'there is a house') are not.

Il y a quelques siècles, les empereurs pourraient modifier les règles selon leur volonté.

```
getUnits(test.sents[["a.65.1"]],
         what = "tokens",
         units = c("DEP_CLAUSES"),
         paste.tokens = TRUE)
```

- Il y a quelques siècles

Ensuite, il y a des délits dus aux problèmes personnels survenus à la maison ou à l'école.

```
getUnits(test.sents[["a.57.1"]],
         what = "tokens",
         units = c("DEP_CLAUSES"),
         paste.tokens = TRUE)
```

- NA

Direct interrogatives in the form *est-ce que* are not considered the head of subordinate clauses. Rather, the head of a clause is the finite verb dominated by *est* as in interrogatives formed by inversion.

Est-ce que les règles sont nécessaires?

```
getUnits(test.sents[["a.17.1"]],
         what = "tokens",
         units = c("CLAUSES"),
         paste.tokens = TRUE)
```

- Est -ce que les règles sont nécessaires

```
getUnits(test.sents[["a.17.1"]],
         what = "tokens",
         units = c("DEP_CLAUSES"),
         paste.tokens = TRUE)
```

- NA

Citations or reported speech enclosed with French guillemets («»), single or double quotation marks are also considered subordinate clauses.

Il est possible que la langue disparaisse après quelques générations,

car, comme dit le professeur Roegiest, professeur des langues romanes et linguiste de l'espagnol à l'université de Gand, « Une langue qui a perdu son prestige est souvent condamnée à mort ».

```
getUnits(test.sents[["b.274.1"]],
         what = "tokens",
         units = c("DEP_CLAUSES"),
         paste.tokens = TRUE)
```

- que la langue disparaisse après quelques générations
- comme dit le professeur Roegiest professeur des langues romanes et linguiste de l' espagnol à l' université de Gand Une langue qui a perdu son prestige
- Une langue qui a perdu son prestige est souvent condamnée à mort
- qui a perdu son prestige

Coordinated clauses

Coordinated clauses are clauses that are not semantically and/or structurally dependent on a super-ordinate syntactic structure but are conjoined to one or more clauses of syntactically equal status. They may be joined by a coordinating conjunction (e.g., *et*; 'and'), punctuation (e.g., semi-colon, colon, comma) or by juxtaposition and must contain both a subject and a finite verb. They are extracted from the root nodes of finite verbs which are not themselves dependent on subordinate conjunctions or pronouns (to exclude dependent clauses).

Dans la prison, il n'ont plus d'éducation, il ne voient que des criminels et ils doivent devenir des adultes en nulle temps.

```
getUnits(test.sents[["a.90.1"]],
         what = "tokens",
         units = c("CO_CLAUSES"),
         paste.tokens = TRUE)
```

- Dans la prison il n' ont plus d' éducation
- il ne voient que des criminels
- et ils doivent devenir des adultes en nulle temps

T-units

We use Hunt's (1970: 199) definition of a T-unit as "one main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it." Identifying T-units therefore depends on the identification of coordinated clauses since a sentence can only contain multiple T-units if it contains multiple coordinated clauses. A sentence that does not contain any coordinated clauses simply has one T-unit, provided it has at least one finite verb. Consistent with Hunt (1965), we do not classify sentence fragments (clauses without a finite verb) as T-units. Therefore, if a sentence has no coordinated clauses (and one finite verb) it has one T-unit. All additional coordinated clauses within a sentence are separate T-units. The three coordinated clauses in the sentence below therefore account for three T-units.

Dans la prison, il n'ont plus d'éducation, il ne voient que des criminels et ils doivent devenir des adultes en nulle temps.

```
getUnits(test.sents[["a.90.1"]],
         what = "number",
         units = c("CO_CLAUSES","T_UNITS"))
```

```
## CO_CLAUSES    T_UNITS
##          3          3
```

Noun phrases

We use Lu's (2010) definition of a noun phrase as a *complex nominal* (see Cooper, 1976) which includes: nouns plus adjective(s), possessive(s), prepositional phrase(s), relative clause(s), participle(s), or appositive(s), nominal clause(s). We also include words (nouns, adverbs and pronouns) that are have a determiner (e.g., *une maison*, 'a house'; *cet autre*, 'this other') in our definition. Following Lu (2010) and Cooper (1976), we also include gerunds and infinitives in subject position. The root nodes of noun phrases therefore include (within each T-unit) common nouns and proper nouns as well as gerunds and non-finite verbs when they are dependent on a finite verb and have a subject dependency label.

La question sur une nouvelle réforme de l'État est un grand problème qui se pose aujourd'hui en Belgique.

```
getUnits(test.sents[["b.46.1"]],
         what = "tokens",
         units = c("NOUN_PHRASES"),
         paste.tokens = TRUE)
```

- La question sur une nouvelle réforme de l' État
- une nouvelle réforme de l' État
- l' État
- un grand problème qui se pose aujourd' hui en Belgique
- La question
- une nouvelle réforme
- un grand problème

Verb phrases

As in Lu (2010) we count both finite and non-finite verb phrases. These are extracted by taking all words which are dependent on a verb within a t-unit. Auxiliary verbs do not constitute their own verb phrase but are considered part of the main verb they modify. However, verb phrases with modal verbs are considered separate verb phrases. After extracting the dependents of each verb node, the units are cleaned by removing subjects and pre-verbal modifiers. When two verbs are coordinated, they are also considered a singular verb phrase (e.g., *ne sont pas d'accord et présentent de solutions différents*; 'do not agree and present different solutions').

Dotée d'un éventail de mots et d'expressions, la parole constitue le moyen de communication par excellence.

```
getUnits(test.sents[["b.124.1"]],
         what = "tokens",
         units = c("VERB_PHRASES"),
         paste.tokens = TRUE)
```

- Dotée d' un éventail de mots et d' expressions
- constitue le moyen de communication par excellence

Appendix II

Table 7.10: Descriptive statistics for sampled LCF texts

		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)

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

Table 7.11: Descriptive statistics for sampled LCF texts (cnt'd)

		Mean (SD)		
	Variable	B2 (n=26)	C1 (n=26)	C2 (n=26)
Syntactic	MLS*	18.04(3.79)	19.06(3.91)	19.01(2.39)
Complexity	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	ADVMOD_V_RTTR	4.54(0.84)	4.51(0.78)	4.86(0.72)
Diversity	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	ADVMOD_V_MEAN_PMI	0.29(0.22)	0.4(0.28)	0.44(0.25)
Sophistication	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)

Appendix III

Definitions of phraseological units in LANGSNAP Study

Adjectival Modifiers (amod)

Adjectival modifiers (amod) were defined as adjectives that modify a noun. This also includes superlatives (e.g., la *drogue* la plus *forte*, 'the strongest drug'). Following Abeillé and Clément (2003), past participles in the attribute position are considered adjectives (e.g., *enfants adoptés*, 'adopted children') except if they are followed by an agentive complement (e.g., *les enfants adoptés par Jean et Luc*, 'the children adopted by Jean and 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*, 'the existing errors'). Quantifiers (e.g., *certaines*, 'certain') and ordinal numbers (e.g., *première*, 'first') are not considered adjectives nor are nouns that are modified by another (proper) noun (e.g., *étudiants Erasmus*, 'Erasmus students'). Compound words (e.g., *fast food*) are not considered adjectival modifiers.

Direct Objects (dobj)

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*, 'the taxes that we pay') or by a passive clause (les *distributeurs* ont été *supprimés*, 'the vending machines were removed') but it does include objects of non-finite verbs (e.g., *concernant* le *mariage*, 'concerning the marriage'). The object of *il y a* ('there is') is considered a direct object relation (e.g., il y *a* une bonne *idée*, 'there is a good idea').

Appendix IV

Descriptive statistics for TEF exams

Table 7.12: Descriptive statistics for TEF data used in model building

	oral, N = 545	written, N = 545
L1		
fra	173 (32%)	173 (32%)
ara	226 (41%)	226 (41%)
hat	11 (2.0%)	11 (2.0%)
kin	12 (2.2%)	12 (2.2%)
other	123 (23%)	123 (23%)
JSD	0.34 (0.17, 0.50)	0.34 (0.17, 0.50)
TOPIC		
.other	422 (77%)	487 (89%)
.travel	123 (23%)	58 (11%)
TOKENS	966 (771, 1,131)	271 (223, 325)
SCORE	551 (499, 598)	475 (404, 549)
LEX.DIVERSITY (MTLD)	40 (35, 46)	58 (50, 66)
LEX.SOPHISTICATION (FRCOWOFF2K)	3.18 (2.49, 3.95)	4.81 (3.55, 6.06)
PHRAS.DIVERSITY (noun-bundles)	0.08 (0.05, 0.12)	0.08 (0.05, 0.13)
PHRAS.DIVERSITY (verb-bundles)	0.07 (0.05, 0.08)	0.07 (0.03, 0.12)
PHRAS.SOPHISTICATION (noun-bundles)	4.68 (4.04, 5.41)	5.91 (5.07, 6.64)
PHRAS.SOPHISTICATION (verb-bundles)	4.81 (4.38, 5.35)	6.22 (5.48, 6.91)

[1] n (%); Median (IQR)

Note:
JSD = Jensen Shannon Divergence (topic difference between oral and written exam sections); MTLD = Measure of Textual Lexical Diversity; FRCOWOFF2K = The number of words in the 3-, 4- and 5-thousand frequency bands of the FRCOW16 corpus (averaged over 100-word moving windows)