Using Machine Learning to Predict Children's Reading Comprehension from Lexical and Syntactic Features Extracted from Spoken and Written Language

by

Jeanne Sinclair

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Department of Curriculum, Teaching, and Learning
University of Toronto

Using Machine Learning to Predict Children's Reading Comprehension from Lexical and Syntactical Features Extracted from Spoken and Written Language

Jeanne Sinclair

Doctor of Philosophy

Department of Curriculum, Teaching, and Learning
University of Toronto

2020

Abstract

Advances in natural language processing (NLP) and machine learning (ML) have introduced exciting prospects to educational research and practice. These technologies are poised to contribute to a deeper understanding of the linguistic and cognitive processes associated with successful reading comprehension, which is a critical aspect of children's educational success. In this thesis, I used ML to investigate and compare associations between children's reading comprehension and 260 linguistic features extracted through NLP from their speech and writing.

Spoken and written language samples were gathered from 172 linguistically diverse children in Grades 4-6 using Talk2Me, Jr., an online language and literacy assessment platform. Lexical and

syntactic linguistic features were extracted via a consolidated NLP pipeline. For the first research question, I compared eight supervised ML models predicting reading comprehension from the linguistic features, and then, using the best model, analyzed the 20 top predicting features. For the second question, I checked for differential functioning by examining interactions between top predictors and language-related demographics in predicting reading comprehension. For the third question, I used unsupervised ML to examine the latent factors constituting the linguistic features and explored how these factors predict reading comprehension differently from the ML models in the first research question. All three parts of the study were performed across four datasets: speech- and writing-elicited linguistic features, for both older/more skilled and younger/less skilled readers.

The study contributes to the literature by concluding that suggest a substantial amount of variance in children's reading comprehension can be predicted by productive grammar and vocabulary. A broad implication is that features of both spoken and written language features correlate with successful reading comprehension, but relationships differ whether individual features or multi-feature factors are used, and whether the data pertain to older/more skilled or younger/less skilled readers. There is evidence that some linguistic features may interact with language-related demographics in predicting reading comprehension, suggesting the need for further research. The study highlights how NLP and ML can enable nuanced examination of the language processes associated with reading comprehension and support innovations in language and literacy research and practice, but also that limitations exist and must be considered.

Acknowledgments

The graduate school journey cannot be traversed alone, and I have many to thank for helping me find my way. Words cannot express my gratitude for the mentorship of my brilliant supervisor, Dr. Eunice Eunhee Jang. During the five years I studied with her, my learning has been, quite literally, immeasurable. Not only did she hold me to higher scholarly standards than I thought I could ever reach, she always insisted that our work be student- and teacher-centered — that it remain applicable to real learning in real schools. I am forever grateful for the multitude of incredible learning opportunities she has provided, for generously sharing her innovative thinking, and for her wise guidance when I most needed it. Dr. Jang's supervision sets a wonderful model for me to emulate. I will always remember how she considers her students' needs and goals first and foremost. I look greatly forward to our continued collaboration.

I am extremely thankful to my committee member Dr. Frank Rudzicz for being instrumental in facilitating my learning of the methodologies employed in this study. Without his support, it would have been nearly impossible for me to navigate the unknown terrain of natural language processing and machine learning. Dr. Rudzicz kindly connected me with Chloé Pou-Prom and Daniyal Liaqat, and all three patiently answered my countless questions. I aspire toward Dr. Rudzicz's impressive commitment to his students' learning and to maintaining high ethical standards in the application of advanced technologies to human lives.

Dr. Jim Hewitt, who also served on my committee, provided excellent and useful feedback on my thesis, and Dr. Earl Woodruff, who served as my internal external examiner, supported me in myriad ways while I was an OISE student. I am grateful for their thoughtful questions, which have helped me consider the implications of this work for teaching and learning, as well as possibilities for meaningful next steps. I also thank my external examiner, Dr. Paula Winke, who provided insightful feedback and thoughtful questions that helped to improve this study.

This study would not have been possible without the teachers and students who participated. I am grateful that participating teachers sufficiently value research such that they are willing to take time from their busy schedules to join this study. Further, the study would not be possible without the students who contributed their thinking, listening, speaking, reading, and writing efforts. Working with them was a privilege and an inspiration, as well as a consistent reminder that the quality of this work hinges upon its utility for teachers and students.

I have been fortunate to receive funding from several sources while pursuing my PhD. I gratefully acknowledge the contributions from the Province of Ontario, the University of Toronto, and OISE, via the Ontario Graduate Scholarship, OISE Graduate Students' Conference Travel Program, OISE Fellowship, OISE Graduate Students' Association, OISE Dean's Office Student Travel Grant, OISE/University of Toronto Graduate Funding, and the University of Toronto School of Graduate Studies Conference Funding program. I am also appreciative for funding contributed by the Learning Environments Across Disciplines partnership in the form of the Student Travel Award.

When I started my PhD journey in 2014, Ontario was a new and foreign land. I very much appreciate Dr. Julie Kerekes taking me under her wing during my first year, and I thank her for the opportunities to collaborate. I am also thankful for the many conversations I had over the years with Dr. Monique Herbert, Dr. Katyn Chmielewski, Dr. Eve Haque, Dr. Saskia Stille, and Dr. Sharla Peltier, who all shared their time and insights with me, providing me with models of the good that academics can do and forming me into the academic I am today. My experiences working with the OISE/Jackman Institute for Child Study faculty, including Dr. Yiola Cleovoulou, Michelle Drimmie Miller, and Anna Totten, further crystalized my commitment to student-centered research and teaching, and I am indebted to them for their apprenticeship.

During my time as a PhD student, I was fortunate to collaborate with scholars across the continent, and their impact on my growth will never be forgotten. Working with the Learning Environments Across Disciplines group was particularly formative, and I would like to thank Dr. Roger Azevedo, Dr. Susanne Lajoie, Dr. Michelle Taub, and Nicholas Mudrick for the wonderfully collaborative research experiences. I am also quite grateful to Dr. Sidney D'Mello and Stephen Hutt for supporting me to participate in the Virtual Learning Lab's exciting research initiatives. Special thanks go to Dr. Charles Lang and Charlotte Woo for creating avenues for us to explore new methodologies in machine learning. I also gratefully acknowledge the contributions of my longstanding mentor, Dr. Francis Hult, who has provided me with thoughtful guidance since the early months of my master's degree in 2009.

My friends and student colleagues in Dr. Jang's IDELA Lab epitomize kindness, generosity, and academic excellence, and I cannot wait to see where their paths lead: Elizabeth Jean Larson,

Megan Vincett, Hyunah Kim, Clarissa Hin-Hei Lau, Christine Barron, Samantha McCormick, Bruce Russell, Melissa Hunte, Liam Hannah, Gina Park, as well as my OISE peers Dr. Shakina Rajendram, Larisa Lam, Stephanie Buono, Klaudia Krenca, Mayo Kawaguchi, and Heba Elsherief. I am eternally grateful for their friendship over the years.

My family has provided me with endless support throughout my time as a graduate student. My wonderful children, Marie, Etta, and Adras, inspire me and spur me to continue forward. I thank my parents and their partners, Mary Helen Dollenmaier and Fred Nadon, and Dr. Michael Sinclair and Susan Woods Sinclair, for setting examples of excellence in everything they do. I cannot sufficiently thank my wonderful husband, Christopher Cessac, for the infinite energy, support, and kindness he has provided our children and me throughout my years of graduate study. He sets a truly impressive example of love, patience, encouragement – and fun – for us all.

Table of Contents

List of Tables

List of Figures

1 Introduction

In recent decades, the use of artificial intelligence has increased in virtually all areas of society. Broadly speaking, artificial intelligence (AI) refers to the study and development of computer systems that execute tasks which are generally considered to be better performed by humans than machines (Rich, 1985). Prevalent examples include navigation, driving, and smartphone technologies, while emerging research focuses on machine vision (training computers to recognize images), conversational AI (focusing on "natural" dialog between humans and computers), and reinforcement learning (teaching computers how to problem-solve and win at games).

In the field of education, AI has focused on two core areas: teaching and learning, and assessment and feedback. Examples of recent research on AI in classrooms includes automatically assessing the authenticity of teachers' questioning – that is, to determine if the questions that classroom teachers typically pose have a preconceived answer or not (Kelly, Olney, Donnelly, Nystrand, & D'Mello, 2018), and applying machine vision to monitor preschoolers' social functioning for mental health intervention (Walczak, Fasching, Cullen, Morellas, & Papanikolopoulos, 2018). In the area of literacy, computer systems that use AI can now recommend reading materials to high school students based on their interests and reading skill (Hsu, Hwang, & Chang, 2010) and provide constructive feedback to young children as they practice cursive handwriting (Simonnet, Girard, Anquetil, Renault, & Thomas, 2019).

Natural language processing and machine learning are analytical techniques that fall under the umbrella of AI, in that they enable the development of computer systems that can learn to complete complex tasks. Natural language processing (NLP) is the programming of computers to interpret and generate human language. NLP allows smartphones to "listen" to commands and "respond", and customer service agents to "chat" with customers. Machine learning (ML) refers to the development of algorithms that can "learn" patterns in complex and large data and then apply that "learning" to novel datasets. For example, ML is the process by which a computer can learn games like chess or Go: the computer determines what moves are more likely to result in success, given the moves leading up to the current state of the game and the system's acquired game knowledge over many iterations of the game.

ML and NLP are often employed in tandem, with ML supporting more efficient NLP (i.e., the computer can correct its errors in understanding and producing human language), while NLP data is a very common raw material used by ML algorithms that seek to predict and understand human behavior. NLP and ML have made some inroads in the field of language and literacy learning and assessment – primarily in automating the assessment of oral reading fluency (e.g., Black, Tepperman, Lee, & Narayanan, 2009), writing (e.g., Burstein, Chodorow, & Leacock, 2004; Crossley, Kyle, & McNamara, 2016), and spontaneous speech (Evanini, Heilman, Wang, & Blanchard, 2015; Zechner, Higgins, Xi, & Williamson, 2009). In the field of literacy, but outside the realm of NLP, ML analytical techniques have been applied to the study of relationships between eye movements and dyslexia risk (Benfatto et al., 2016).

However, with regard to reading comprehension and its associated skills, arguably the topmost concern of the educational enterprise, the technological power of NLP and ML has not been adequately leveraged. The present study explores the potential, and limits, of these advanced technologies for research on, and assessment of, reading comprehension and the linguistic processes enabling it. In this study, I engage with decades-strong research on the relationship between reading and broader language processes. The Simple View of Reading (Hoover & Gough, 1990) – the most well researched and widely debated theory of reading comprehension in contemporary research – suggests there is a strong relationship between reading comprehension and language comprehension, or the ability to make meaning from the language one hears. Hoover and Gough's original study was developed with bilingual (Spanish-English) elementary students, with a focus on reading comprehension in English. The authors found that two fundamental constructs enable reading comprehension: first, phonological decoding, which consists of converting printed symbols of text to phonemes (the fundamental sound units that comprise words), and then blending the phonemes to make words, and second, language comprehension, which is also known as linguistic or listening comprehension (Hogan, Adlof, & Alonzo, 2014). Hoover and Gough's seminal work was introduced at a time when two schools of reading research warred over the primacy of phonics- or whole language-oriented approaches to reading instruction. The Simple View was an attempt to reconcile these by situating reading comprehension as a product of both.

Subsequent studies have deconstructed the language comprehension construct in a variety of ways, with the goal of finding the specific contributions of different combinations of language

comprehension's component constructs, including working memory, attention, vocabulary, grammar, morphology, inferencing ability, activation of relevant background knowledge, and metacognition and monitoring (Cain & Oakhill, 2014; Deacon & Kieffer, 2018; Hagtvet, 2003; Kim, 2017; Kim, Park, & Park, 2015; Lesaux, Lipka, & Siegel, 2006; Nation & Snowling, 2004; Ouellette & Beers, 2010; Poulsen & Gravgaard, 2016; Tannenbaum, Torgesen, & Wagner, 2006; Tunmer & Chapman, 2012). Another important line of research found a strong predictive relationship between reading comprehension and productive oral language, beyond receptive oral language comprehension (Catts, Fey, Zhang, & Tomblin, 1999; Foorman, Koon, Petscher, Mitchell, & Truckenmiller, 2015; Holahan et al., 2018; Kendeou, Bohn-Gettler, White, & van den Broek, 2008; Kim et al., 2015, Nation, Clarke, Marshall, & Durand, 2004; Scarborough, 2005). Further, the relationship between writing quality and reading comprehension has been well documented (e.g., Berninger & Abbott, 2010; Berninger, Abbott, Abbott, Graham, & Richards, 2002; Carretti, Motta, & Re, 2016; Shanahan, 2016). In sum, extensive research has found interrelationships between productive and receptive language, in both spoken and written domains: speech, listening, writing, and reading.

Two primary constructs underlying all language domains are lexis (the depth and range of the vocabulary one can understand and produce) and syntax (the complexity and accuracy of the grammar one can understand and produce). These constructs have been studied extensively in both spoken and written formats to understand how much variance in reading comprehension can be attributed to them (e.g., Brimo, Apel, & Fountain, 2017; Cain, 2007; Gottardo, Mirza, Koh, Ferreira, & Javier, 2018). However, many studies in this area have focused on receptive measures (one-word answer or selected-response item formats), while few have used productive measures; that is, relatively fewer studies have examined the relationship between reading comprehension and the lexical and syntactic features of open-ended spoken and written responses. This is in part due to the relative ease of scoring receptive measures. Yet, assessing language and literacy skills through productive, constructed language responses elicits deeper cognitive processing and perhaps more information than selected responses (Pearson & Hamm, 2005). Thus, there is a pressing need to examine how advanced technologies can facilitate the assessment of productive, constructed language.

In the present study, I used NLP and ML techniques to examine the relationship between children's reading comprehension and the grammar and vocabulary of their productive language.

I compare this relationship for oral and written domains to determine commonalities and differences. In other words, does the vocabulary and grammar in children's speech predict the same variance in reading comprehension as the vocabulary and grammar they use when writing? The study seeks to understand the broader language skill sets associated with reading ability, and how NLP and ML may support a finer-grained understanding of that association, beyond that offered by traditional measures and analytical techniques. To my knowledge, this is the first study to use NLP and ML for this purpose.

The ML models in this study take two forms: supervised and unsupervised. In the supervised approach, the algorithm is trained to recognized patterns in a given dataset using known classifications or scores; these are called "labels." Then, if pertinent to the research question, the variables that best predict those labels can be identified from among the variables in that dataset, which is known as "feature selection". As a hypothetical illustration, a supervised ML modeling approach could be used to predict students dropping out of school from a set of data that includes grades, attendance, and participation in extracurricular activities. If the research question calls for such specification, it is possible to examine the functioning of individual variables within the model in order to identify which variables are most predictive of school drop-out. Using a training set, or portion, of the dataset, the ML algorithm identifies which variables are associated with the school-leaving outcome (label), including the strength and direction of the relationship. Then, the remaining portion of the dataset (the "test" set) is used to evaluate how well the algorithm learned the relationship: the algorithm reads in the test set's predictor variables, and, based on the previous learning during training, attempts to classify the observations as school-leavers or not-school-leavers. Then, the accuracy of those predictions is evaluated by comparing the algorithm's predicted outcomes with the true outcomes. The greater proportion of true outcomes that the algorithm predicted correctly, the better the accuracy of the algorithm. The unsupervised approach, on the other hand, clusters, classifies, or orders data without known labels. Akin to factor analysis, clustering, and principle component analysis, patterns in the predictor dataset are identified. The unsupervised ML approach is commonly used to reduce the dimensionality of wide datasets.

Three studies comprise this thesis. For each, I used four datasets following a two-by-two research design matrix: samples of *more* and *less* skilled readers, and, for each sample, productive language features were elicited through writing (*text-based*), and through spontaneous

speech (*oral-based*). The same lexical and syntactic features were extracted for each of these four samples. The three research questions (RQs) are described below.

RQ1. How much variance in reading comprehension can be modelled as a function of productive lexical and syntactic features extracted through NLP? How does the variance explained compare to that explained by models that use traditional lexis and syntax measures? What are the top lexical and syntactic feature predictors for each of the four models (oral/text elicitation by more/less skilled readers)? How do the top predictors differ across these models?

In this study, I explore what proportion of variance in children's reading comprehension can be modelled as a function of individual NLP-derived lexical and syntactic features. The goal of this analysis is to determine whether a similar amount of variance in my participants' reading comprehension score can be explained with NLP-derived lexical and syntactic features as with traditional measures of lexis and syntax. The extant literature suggests lexical and syntactic factors explain between 25% (Poulsen & Gravgaard, 2016) to 96% of variance in reading comprehension (Foorman et al., 2015), when those factors are measured by traditional standardized or human-scored measures.

Each of the four lexical and syntactic feature datasets (*more* and *less* skilled readers by *speech* or *text* elicitation) is wide (having more variables than observations), thus precluding the use of traditional multiple regression when predicting the reading comprehension score. For this reason, and because I cannot assume linear relationships among the data, I used ML for model-building. ML also offers an advantage of the capability to test the model's predictive generalizability, by testing its predictive power on an unseen portion of the dataset. For each of the four datasets, I compared eight different ML approaches against a mean baseline to determine the best fit based on mean absolute error. Once the best ML model was chosen, I analyzed and compared the top lexical and syntactic feature predictors for oral-elicited and text-elicited datasets. I hypothesized that text-elicited lexical and syntactic features will explain more reading comprehension variance than features elicited from speech. I also predicted that similar lexical and syntactic features will be strong predictors in both speech- and text-elicited data, although top predictors may differ in the more and less skilled reading groups. Restated, I expected more variation in the nature of the

top lexical and syntactic feature predictors across the two reading skill groups (since they are two different populations) than depending on how the features were elicited (speech or text).

RQ2. Is there significant interaction between the top lexical and syntactic features and students' length of residence in Canada, or their multilingual proficiency in languages other than English, in predicting reading comprehension? If interactions exist, how do they differ across the four models?

In the second part of this study, I examined whether the top predictors in each model function differently depending on immigration background and multilingual proficiency. A relatively large body of research has examined differential functioning of language and literacy assessment items for students of diverse language and immigration backgrounds. Although the use of NLP in language assessment has enormous potential for improving teaching and learning, very little is known about how the strength of the relationship between NLP-derived linguistic features and the outcome variable may differ for different demographic groups. I performed tests for interaction (akin to differential item functioning) between demographic group and the NLP-derived features in predicting reading comprehension. Specifically, I focused the interaction analyses on two variables: multilingual proficiency, and whether or not participants were born in Canada. These variables were selected because existing literature has shown differential functioning of traditional language and literacy assessment questions, or items, based on students' language and immigration background (Kim & Jang, 2009; Koo, Becker, & Kim, 2014), meaning that some assessment items favor certain language or immigration backgrounds when overall ability is held constant. If assessment items demonstrate measurement bias against a certain group, then the assessment may not be a fair measure of those students' abilities. In this case, my goal was to determine if the NLP-derived lexical and syntactic features predict reading comprehension with the same strength and direction of relationship for participants who reported being multilingual or born abroad.

RQ3. What latent factors can be identified in the NLP-derived data using an unsupervised ML method? Do the factors differ across the four models? Do the resulting factors predict reading comprehension? How do these predictive relationships differ from those in the first research question?

In the third part of this study, an unsupervised ML approach was performed to understand the underlying structure of the NLP-derived syntactic and lexical features. The resulting factors were entered into regression equations to determine if any had a predictive relationship with reading comprehension. Then, the results were compared with the supervised methods to determine commonalities and areas of difference – do the factors underlying the NLP lexical and syntactic features predict reading comprehension similarly to how individual NLP features do?

1.1 Significance of the study

This thesis contributes to both theoretical and methodological issues in language and literacy research. In terms of theory, an extensive body of research has examined how receptive language skills predict reading comprehension. I explore the predictive relationship between reading comprehension and productive (written and spoken) vocabulary and grammar skills, as well as how this relationship may differ for more or less skilled readers. Methodologically, the thesis is novel in its use of fine-grained linguistic features derived through natural language processing to address the question of how lexical and syntactic language skills relate to reading comprehension (RQ1). Another methodological and theoretical contribution is understanding the factor structure of the natural language processed lexical and syntactic features, and how those factors may predict reading comprehension similarly or differently to how individual features predict reading comprehension (RQ3). Essentially, can productive language features predict reading comprehension as well as receptive language features? Does the use of fine-grained NLP-derived language features offer any advantage in terms of understanding how lexical and syntactic skills relate to reading comprehension? Finally, the study is novel in its investigation of how variables relating to immigration background and multilingual ability may interact with the NLP-derived lexical and syntactic features in predicting reading comprehension (RQ2), which is important because these results impact the generalizability of the findings of RQ1.

1.2 Overview of the thesis

Chapter 2 reviews the literature on the four domains of language (reading, writing, speaking, and listening), with a focus on how the latter three relate to reading and the role of syntax and lexis. Chapter 3 provides a detailed description of the NLP and ML techniques used in the three studies. Chapter 4 delineates the findings, and Chapter 5 provides a discussion for each study as

well as a general discussion of how the three studies' findings relate to one another. Finally, Chapter 6 presents the study's limitations and conclusion.

2 Literature Review

The literature review first discusses the research base on theoretical issues of language and literacy, specifically the interrelation between the language domains, the processes that enable reading comprehension, and research relating to assessment and students from diverse linguistic and cultural backgrounds. This is followed by a focus on the methodologies employed in this thesis, especially natural language processing and machine learning.

2.1 Language and literacy: interwoven skills

This section describes the literature on the interrelationships between the four domains of language: reading, writing, listening and speaking, with a focus on how the latter three relate to reading comprehension. Two core constructs of language – syntax and lexis – were highlighted as predictors of reading comprehension, as well as NLP tools for their analysis. Syntax and lexis have shown to be essential factors in all four language domains. While the majority of research operationalizing syntax and lexis as part of the oral language construct has used receptive measures, the present study investigates their role in productive oral language, as well as productive written language.

2.1.1 Relating reading comprehension and oral language

Reading comprehension – the construction of meaning from printed text – is the goal of reading instruction as well as a key predictor of school success (Lesaux, Rupp, & Siegel, 2007). While educational programs and curricula tend to focus on the development of reading comprehension, it is oral language defined more broadly that forms the foundation for reading comprehension. The interrelation between language and literacy has been the subject of research from a wide range of theoretical perspectives, from the multiliteracies paradigm (Cope & Kalantzis, 2000) which explores this interaction in the context of broader social, political, and cultural milieus, to neuroimaging studies examining the nature of language and literacy subskill interactions for children with different language-related disabilities (Krishnan, Watkins, & Bishop, 2016).

Broadly speaking, oral language refers to the ability to communicate verbally through listening and speaking: "all of verbal ability, including vocabulary, syntax, inferencing and the construction of mental schemas…[and it may be] the greatest achievement of human evolution" (Kirby and Savage, 2008, p. 76). According to Dunst, Trivette, Masiello, Roper, and Robyak (2006), oral language refers to a shared code used to communicate ideas: it is a set of rules, including "phonological (the rules for combining sounds), morphologic (rules for the internal organization of words), semantic (word meaning), and syntactic (rules that have to do with the order of words in sentences) elements of language" (p. 4). Yet, despite ample research addressing children's oral language, questions remain regarding exactly what it is, how it can be differentiated from literacy, and how it specifically influences literacy development (Jang, Cummins, Wagner, Stille, & Dunlop, 2015; Lesaux, Geva, Koda, Siegel, & Shanahan, 2008). Kim (2017) argues that language comprehension remains an underspecified "black box" because the relationship between its myriad components is not well understood.

What is known is that oral language plays an important role in the development of reading comprehension from early childhood through adolescence and early adulthood. For preschoolers and kindergarteners, oral language deficits are associated with an elevated risk for developing a reading disability, even if those oral language deficits appear to be remediated in the primary years, as summarized by Scarborough (2005). Catts et al. (1999) retrospectively examined the kindergarten performance of monolingual English-speaking Grade 2 students identified as less skilled readers. They found that 57% of the Grade 2 less skilled readers showed deficits (in kindergarten) in receptive oral language and 50% in expressive oral language, a rate four times greater than among skilled readers. Only 14% of less skilled readers in Grade 2 exhibited solely phonological processing issues without broader oral language deficits in kindergarten. Kendeou, van den Broek, White, and Lynch (2009) tracked monolingual English-speaking preschooler's oral language and decoding skills to Grade 2, and found that oral language (measured as receptive vocabulary and aural/video narrative retelling) and decoding skills formed separate but correlated clusters, with the relationship becoming progressively weaker over time. Both skill clusters independently predicted Grade 2 reading comprehension. Nation et al. (2004) found that 8-year-old children (L1 English users) with normal phonological abilities but poor reading comprehension skills were also likely to have weaknesses in a variety of oral language skills including semantic skills, morphosyntactic skills, and broader language skills (ambiguous

sentence identification, listening comprehension, oral expression, and figurative language tasks). They also found that a substantial proportion of these learners would also qualify as having specific language impairment based on the measures employed. Indeed, language impairments identified during the preschool years can predict future comprehension-related reading exceptionalities (Scarborough, 2005). The connection between oral language and reading is a main focus in the present study.

Comprehension is not confined to comprehending solely what one reads. Comprehension has been shown to be a universal cognitive process operating across both language-based (aural and written) and image-based stimuli, as Gernsbacher, Varner, and Faust (1990) demonstrated with a sample of undergraduates who were L1 English users. With neuroimaging, Braze et al. (2011) and Buchweitz, Mason, Tomitch, and Just (2009) found with a sample of young adults that comprehension of sentences in written and spoken forms activate some of the same regions of the brain, as well as some modality-specific sites. The participants in the former study were L1 English users, while the latter involved L1 Portuguese users.

To comprehend language, be it aural or written, one must create a "mental model": a series of constantly-updated representations that one develops and maintains in the mind as one encounters text (here "text" is used to denote both written and oral language). According to Kintsch's discourse-processing model (1994, 2012; c.f., Kintsch & Kintsch, 2005), comprehension takes place over two phases. First, a propositional model is developed – essentially a skeletal interpretation of the propositions set forth in written or spoken language. The propositional model requires lower-level language processes such as vocabulary, morphology, and syntax. Then, listeners/readers create a mental, or situational, model by integrating their prior knowledge, goals, and even emotions with the propositional model. The process of developing a situational model requires higher-level cognitive processes such as discourse-level language processing, inference generation, and monitoring (Perfetti, Landi, & Oakhill, 2005). Developing a coherent situation model is the goal of comprehension (Kintsch & Kintsch, 2005), because once listeners/readers situate and contextualize the information they hear or read, it can be filed among the appropriate schemata – one's filing system for knowledge – to be accessed later (Cain, 2015; Kintsch, 1998).

Although Kintsch's model was originally applied to reading, it can be applied to oral language comprehension as well (Hogan et al., 2014), since language comprehension (regardless of the medium) is the process of integrating new propositions and relationships into the existing schemata in the mind of the reader/listener (Kim, 2017). Like reading comprehension, language comprehension requires making inferences such as deciding which background knowledge to call into play and which to ignore or prune away (Anderson & Pearson, 1984). Of course, reading comprehension requires additional skills beyond language comprehension, such as decoding accuracy and fluency; oral language comprehension, on the other hand, may require additional working memory resources than reading comprehension since the text is not available to reread (Wolf, Muijselaar, Boonstra, & De Bree, 2018). Nonetheless, the relationship between oral language comprehension and reading comprehension is evident: readers who struggle with language comprehension have difficulty building appropriate mental models of texts, making appropriate inferences, understanding story structure, and resolving anomalies or discrepancies in texts (Oakhill, Cain, & Yuill, 1998).

Reading comprehension requires an interaction of multiple linguistic and cognitive skills that vary from learner to learner (Cain & Oakhill, 2006), with all levels of oral language playing important roles in constructing and maintaining a mental model (Cain, 2015). In a unique longitudinal study of inference-generation across media, Kendeou et al. (2008) followed two cohorts of children: one from ages 4 to 6 and the other from ages 6 to 8 (language background was not included in this study). The authors analyzed the children's inference skills using audio, video, and written stimuli. They found that the skill of inferring is generalizable across audio/video and written media types, and that this skill is unrelated to sound-symbol decoding skills such as phonemic awareness and letter and word identification. The authors also found that this media-independent inference-generation skill increases with age and is the strongest predictor of overall reading comprehension of all constructs measured in the study.

With participants who were second language learners of English from Spanish-speaking backgrounds, ages of 9 and 13 years, Gottardo et al. (2018) also relate oral language comprehension to reading comprehension. These authors emphasize that the construct of language (listening) comprehension contains crucial cognitive and linguistic components: "knowledge of vocabulary, morphology, and syntax… working memory, allocation of attentional resources, higher order reasoning skills… [and] general and topic-specific background

knowledge" (p. 1743). In their study, a language comprehension construct including vocabulary, morphology (measured using a derivational awareness task), and syntax explained 67% of the variance in reading comprehension for preadolescents learning English as an additional language (EAL).

This framework of unitary language comprehension across reading and listening domains underlies the "simple view of reading" (SVR; Gough, Hoover, & Peterson, 1996; Hoover & Gough, 1990; Gough & Tunmer, 1986). The SVR hypothesizes that reading comprehension (RC) is a product of the interaction between decoding (DC) and language comprehension (LC) skills, denoted as RC = DC x LC. Decoding refers to the ability to apply phonological awareness (that is, the knowledge that words are made of phonemes, or sounds, and the skill to manipulate phonemes to make new words) and knowledge of letter-phoneme relationships to identify the written symbols of language and thus read written words. When language comprehension and decoding are both well developed, then reading comprehension is likely to be well developed, too. According to the model, if either language comprehension or decoding are less developed, this will have a multiplicative effect on reading comprehension ability. Hoover and Gough proposed a multiplicative model because an additive model assumes that if decoding is at zero then reading comprehension would still be possible. For example, in theory, if decoding skill is zero, then reading comprehension must also be zero. The SVR is depicted in Figure 1.

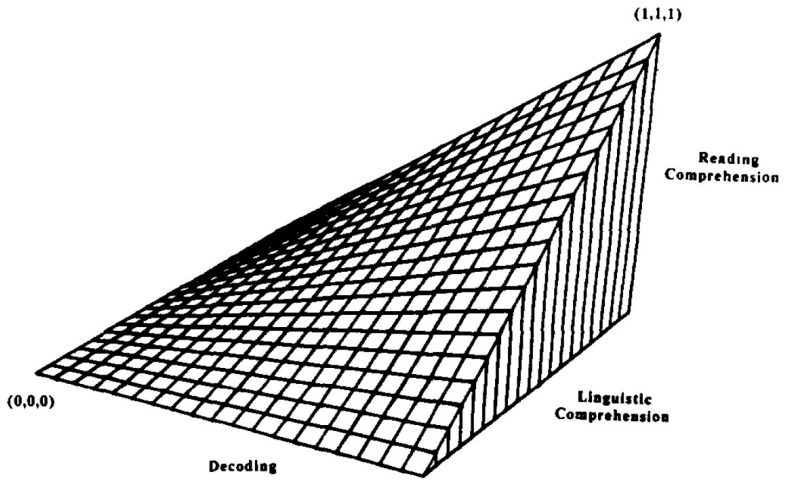

Figure 1. Model of the Simple View of Reading (Hoover & Gough, 1990)

Gough and Tumner (1986) summarize the SVR as postulating that "once the printed matter is decoded, the reader applies to the text exactly the same mechanisms which he or she would bring to bear on its spoken equivalent….It would be falsified [by] someone who could decode and listen, yet could not read" (p. 9). The SVR as originally conceived considers listening comprehension rather than broader oral language capabilities. While the present study is

interested in the relationship between productive oral language and reading comprehension, the SVR is the most significant contemporary theory of reading that informs research on the relationship between reading comprehension and broader language processes. Thus, while focusing on different angles, the SVR is a touchstone for the studies discussed in the remainder of this section.

Kim (2017) sought to understand the skills underlying the language comprehension component of the SVR by testing different structural relationships between cognitive and linguistic factors with a sample of Grade 2 students (98% of whom were L1 English users). She found a hierarchical model fit best: attentional resources and working memory are foundational cognitive skills, and with them the foundational language skills of vocabulary and grammar can be used to access and produce higher-order cognitive processes such as inference generation, metacognition, and perspective-taking. Kim's hierarchical factor structure explains 86% of variance in discourse-level language comprehension and 66% of variance in reading comprehension. The best-fitting model that also included a decoding variable explained nearly all the variance in reading comprehension, with word-reading and language comprehension completely mediating the relationship between reading comprehension and all component language and cognitive skills; these results can be interpreted as validating the SVR.

Deficits in language comprehension become more pronounced and have a greater impact on reading comprehension as learners mature and the texts and cognitive demands of the curriculum become more challenging (Catts, Hogan, & Adlof, 2005; Tilstra, McMaster, Van den Broek, Kendeou, & Rapp, 2009). Nation and Snowling (2004) assessed L1 English-speaking children at ages 8.5 and 13 and found that the influence of language comprehension contributed 31% of total variance at age 8, when entered as the last step of a hierarchical regression, above and beyond nonverbal ability, nonword reading, phonological skills, semantic skills, and vocabulary. At age 13, language comprehension measured five years prior explained 14% of unique variance in reading comprehension, even after all previous factors were included in addition to the autoregressive timepoint 1 reading comprehension score. The increased share of variance explained by language comprehension can be attributed to the increased complexity of texts as students progress through school: by adolescence, decoding no longer plays an important role for most readers, but texts, even when read aloud, can be challenging to understand.

Indeed, oral language ability also remains important in secondary school, where it has been shown to be a strong predictor of high school academic achievement (in both English and math), even exerting a stronger influence than socioeconomic status (Spencer, Clegg, Stackhouse, & Rush, 2017); this research sample included 13- and 14-year-old children, 24% of whom used more than one language at home. Given that language comprehension is a major component of oral language, this makes intuitive sense, because language comprehension predicts a greater amount of variance in reading comprehension as learners progress from elementary to high school (Catts et al., 2005). In a longitudinal study of oral language, reading fluency, and reading comprehension across L1 English-using identical twins aged 7-16 (Tosto et al., 2017), individual differences in reading comprehension were largely explained by genetics at age 7, but the genetic (that is, non-environmental) influence on oral language rose further as children approached adolescence. The authors also found that both genetic heritability and environment play a role in oral language development from middle childhood through adolescence. A longitudinal study of the relationship between oral language and reading comprehension for L1 English users in grades 1 through 9 (Holahan et al., 2018) found that all components significantly predicted the reading comprehension outcome measure, with long-term background knowledge and vocabulary showing the strongest correlations.

While the title "Simple View of Reading" suggests this is a simple theory, the SVR is not as simple as the name suggests. Accurate and efficient decoding is developed through the acquisition of phonological awareness, a process which can take years to develop. Efficient decoding is also associated with reading fluency – the speed, accuracy, and prosody with which a learner reads – which is another key reading construct. Silverman, Speece, Harring, and Ritchey (2013) and Tilstra et al. (2009) have demonstrated in samples of L1 English users in grades 4, 7, and 9 that reading fluency is an additional construct in the SVR; however, Adlof, Catts, and Little (2006) have shown in a study of children in grades 2-8 that fluency does not explain additional variance beyond decoding and listening comprehension.

Other modifications to the SVR have been proposed. Chen and Vellutino (1997) cross-validated the SVR with a sample of students in Grades 2,3,6, and 7, and found that including an additive component along with the multiplicative component better explains the relationship between reading comprehension, listening comprehension, and decoding. This conclusion is based on their finding that the models' intercepts predicting reading comprehension from listening

comprehension were nonzero. They found that listening comprehension and decoding predicted 58% of the variance in reading comprehension in Grade 2, 61% in Grade 3, 65% in Grade 6, and 58% in Grade 7, using the combined additive/multiplicative model. Chen and Vellutino's proposed "weak" SVR model relaxes the strict assumption of independence between listening comprehension and decoding, which they argue is necessary because listening comprehension predicts reading comprehension differently depending on decoding skill, with slopes increasing at higher levels of decoding. In other words, decoding has a relationship with listening comprehension in that it acts as a moderator between it and reading comprehension.

Another important interpretative update on the SVR is Stanovich and Siegel's (1994) seminal study comparing the phonological, oral language, and cognitive skills of reading-level matched children aged 7-16 with and without substantial reading difficulties; all were L1 English users. Participating children with a reading disability were divided into two definitional groups: one group based on a discrepancy definition (scoring below the 25th percentile in word reading, and having average to above-average standardized IQ score), and those without a discrepancy (like the previous group, scoring below the 25th percentile on a word reading measure, but also having scored below-average on the IQ measure). Participants varied in age because they were matched by reading level. The authors found that while the discrepancy and non-discrepancy defined groups differed in working memory, the two groups did not differ significantly on most language skills. Scarborough (2005) concurs, proposing the possibility that rather than being causally implicated in reading disability, phonological awareness itself is but one presentation of broader oral language factors that may comprise "the condition we call reading disability [arising] most fundamentally from an underlying…predisposition to process complex verbal material less efficiently" (p. 17). This suggests that that deficits in broader oral language skills (which include language comprehension) are implicated in phonological deficits, which in turn underpin decoding deficits, providing further evidence for the importance of oral language development as a prerequisite for successful reading comprehension.

Hagtvet (2003) examined the influence of vocabulary, syntax, phonemic awareness, working memory, and IQ in predicting comprehension (measured via story retelling and cloze tasks) of oral and written texts for poor, average, and good decoders (aged 9). All participants were L1 users of Norwegian. Concurring with Stanovich and Siegel's (1994) findings, poor decoders also had the lowest scores on the oral language measures, despite her study's inclusion criteria of no

manifest oral language delays and average to above-average intelligence. Thus, Hagtvet also concludes that the broad set of oral language skills that enable reading comprehension also include phonological awareness and decoding ability. This finding suggests there may not be qualitative differences between readers who are strong aural comprehenders but struggle due with decoding (who are typically identified as having dyslexia) and the so-called "garden variety" less skilled readers who have difficulties with language comprehension and decoding. Instead, these two learner archetypes may differ along a continuum of language development. Like Chen and Vellutino's proposed weaker model (1997), Hagtvet's findings contradict Hoover & Gough's "strong" model, wherein language comprehension and decoding are independent, and instead allow decoding and listening comprehension to interact.

Hagtvet's (2003) study was novel in that all measures were administered through both oral and written versions (with different wording of actual items to avoid practice effects). Supporting the idea that aural and written language skills have strong intraindividual relationships, all correlations between the oral and written versions of each measure were significant at $p<.01$. She notes that her team was surprised to find that

> … listening and reading comprehension of equivalent types of tasks would share so many similarities in terms of underlying oral language skills …[This] underscores the importance of task demands, to some extent over and above modality [print or aural]….Skills observed in connection with written comprehension tasks when only written abilities are focused may therefore deceive the researcher to see modality specific skills where there are none. (p. 525-6)

In summary, Stanovich and Siegel (1994), Chen and Vellutino (1997), and Hagtvet (2003) suggest that the SVR's original specification of independence between broader oral language skills and phonological awareness and decoding skills may be falsifiable. While the present study does not address decoding skills per se, these findings lend support to the idea that oral language skill development is of prime importance for children's reading growth. Beyond the constructs typically associated with oral language (vocabulary, syntax, etc.), these studies suggest even decoding and phonological skills can be impacted by broader oral language development. This

has significant implications for the way educational institutions identify and intervene for reading-related difficulties. Currently, "unexpected" deficits in phonological awareness are generally a primary indicator for a reading-related learning disability, which may inadvertently ignore the needs of children who struggle in both general language comprehension and decoding, and therefore for whom deficits in phonological awareness are not "unexpected" – despite evidence that typical interventions for reading-related disabilities are equally effective for both the discrepancy and non-discrepancy groups (Fletcher, Lyon, Fuchs, & Barnes, 2018; Stage, Abbott, Jenkins, & Berninger, 2003; Stanovich, 2005). Thus, the focus on phonological "unexpectedness" (i.e., decoding deficits without the presence of other difficulties) when identifying students for Tier III (individualized) interventions may be misguided, and the presence of oral language deficits in young children may be sufficient to warrant intervention.

Another key area of research that relates to the present study is the relationship between language comprehension and oral language: are they separate constructs, or is language comprehension a component of oral language? The original SVR is based on measures of receptive language comprehension (operationalized as listening comprehension) and receptive reading comprehension. The present study, however, seeks to understand how productive language (in speech and writing) relate to reading comprehension. Thus, the relationship between language comprehension and oral language is an important one to consider. Gray, Catts, Logan, and Pentimonti (2017) examined the factor structure of language comprehension and oral language through grammar, vocabulary, and listening comprehension tasks in a sample of children from pre-kindergarten through Grade 3, with 22% reporting multiple languages spoken at home. While a two-factor model (listening comprehension and oral language) best fit the data, the two factors were correlated at .91 and therefore can be considered to operate as a single factor. Lonigan and Milburn (2017), in a study of typically developing children in pre-kindergarten through Grade 5, and Tomblin and Zhang (2006), whose sample included children in kindergarten through Grade 8 (language background was not provided) also found little evidence for separate expressive and receptive oral language factors. In a study of Grade 1 students, Kim et al. (2015) examined the factor structure of listening comprehension, oral retell, and oral production tasks with L1 Korean users in Grade 1, finding a bifactor structure including an overall discourse-level oral language construct with separate underlying listening comprehension and oral retell constructs. A cross-sectional study of students in grades 4 though

10 (Foorman et al., 2015) found that the vast proportion of variance in reading comprehension is explained by a general oral language factor that includes syntax and vocabulary; the research sample included between 0% and 20% students learning EAL, depending on the grade level.

As for students who are learning EAL, Prevoo, Malda, Mesman, & van Ijzendoorn, (2016) demonstrated through metanalysis that the predictive power of oral language for reading comprehension increases from lower to higher grade levels. For students learning EAL in upper elementary school, oral skills in the second language have a significant impact on reading comprehension in the second language, over and above word-reading skills (Lesaux, Crosson, Kieffer, & Pierce, 2010). Oral language has also been identified as the strongest predictor of reading comprehension in nine-year-old children, for both L1 and L2 English users (Babayiğit, 2015). Babayiğit (2014) researched the relationship between oral language and reading comprehension for upper elementary EAL and non-EAL learners, concluding that EAL learners' oral language (operationalized as vocabulary and morphosyntactic skills) explained variance in their listening and reading comprehension; less variance was explained for non-EAL learners. Both EAL and non-EAL learners who struggle with meaning- related reading skills are likely to sustain those difficulties over time, as comprehension is a more challenging area for intervention than code-related skills such as decoding and fluency (Geva & Herbert, 2013; Lesaux & Harris, 2013; Dickinson et al., 2010). However, comprehension-based reading interventions have been shown to be one of the most effective forms of intervention in the long-term for struggling readers (Suggate, 2016).

2.1.1.1 Issues assessing comprehension and oral language

Comprehension of both audio and written stimuli is a latent, internal process that is notoriously difficult to assess and intervene upon. Further research is still needed about assessment and intervention, especially for language comprehension skills (Cain, 2015; Catts et al., 2005). Hogan et al. (2014) describe 15 different language comprehension assessments employed in research studies since 1989, and the skills elicited by those assessments vary widely. In some assessments, text paragraphs are read aloud, and participants must answer either discrete or open-ended literal or inferential comprehension questions. Other assessments ask participants to follow a command (e.g., to point to a certain picture), with items increasing in grammatical and syntactic complexity. Hogan et al. (2014) conclude that "the field lacks specific

recommendations about how best to assess development in language comprehension or how to intervene when language comprehension skills are not up to par" (p. 200). Gottardo et al. (2018) concur that "… in most studies, the construct of listening comprehension is not well defined" (p. 1743).

With regard to reading comprehension, for decades, scholars have also lamented that reading comprehension assessments can only capture reading products, from which inferences about the nature of processes could be inferred (Pearson & Johnson, 1978). In addition, many contemporary reading comprehension assessments are built on outdated theories of literacy and learning (Pearson, Valencia, & Wixson, 2014; Kintsch & Kintsch, 2005). Assessing reading comprehension through selected-choice questions in a testing environment may elicit reasoning and test-wiseness strategies which are quite different from learners' reading processes in non-testing environments (Rupp, Ferne, & Choi, 2006). Open responses to reading, or reading-to-write responses, are an alternative to selected-choice questions. For example, retelling after reading is a common approach to assessing comprehension; however, retelling may not elicit higher-order thinking processes such as inference-generation which are increasingly important as students progress through school (Shapiro, Fritschmann, Thomas, Hughes, & McDougal, 2014). Generating a summary, where a condensed version of the text is presented, is another form of reading response. The quality of upper elementary and middle-school students' (4% were learning EAL) summaries of nonfiction texts have been shown to be influenced by their comprehension of the text and also the linguistic (lexical, syntactic, and discourse-level) resources they bring to the task, after controlling for student demographic characteristics (Galloway & Uccelli, 2019). Classroom-based reading comprehension assessments include informal one-on-one reading inventories, anecdotal records and observations, retelling and recalling, freewriting, interviews and think-aloud procedures, short-answer questions, and cloze techniques (Fuchs, Fuchs, & Maxwell, 1988; Klingner, 2004). These classroom-based techniques, used in combination, tend toward strong validity but are difficult to utilize in large-scale assessments. The current study explores the possibility for assessing children's productive language in literacy-focused tasks through NLP and ML, which it is hoped can support the improved validity of computer-based language and literacy assessments.

In contemporary research studies, the operationalization of oral language tends to encompass one or more named skills such as receptive and expressive vocabulary, syntactic knowledge,

semantic knowledge, and narrative discourse processes including recall, comprehension, and storytelling (National Institute of Child Health and Human Development Early Child Care Research Network, 2005). There is a need for quality assessments of oral language (Lesaux et al., 2008). Measures purporting to assess oral language have been shown to have different dimensionality than their published manuals claim. In the case of one study of two such measures (Hoffman, Loeb, Brandel, & Gillam, 2011), which involved children ages 6-8 from predominantly English-speaking homes, there was 64% unexplained variance among the two measures when students who were known to have specific language impairments were assessed. Not all oral language assessments are appropriate to use with all populations, and students' cultural and linguistic backgrounds, as well as potential impairments, should be considered (Dockrell & Connelly, 2015). Humphry, Heldsinger, and Dawkins (2017) developed an alternative assessment tool to measure the narrative production of young children during story recall based on teachers' judgment using performance descriptors. The skills assessed included the ability to tell a story, sequencing and cohesiveness of ideas, length of sentences and variety of sentence beginnings used, sentences' grammatical structure of sentences, correct use of tense, vocabulary and descriptive language, and word articulation. These efforts are useful in developing valid tools that capture the enormous construct that is oral language.

Ample research has addressed the importance of oral language development – and language comprehension specifically – for children, as it provides the cognitive and linguistic foundation for successful literacy development. Nonetheless, oral language is not a commonly targeted skill in curricula, and this crevasse between research and practice has significant negative consequences. While the "word gap" or "language gap" (Hart & Risley, 1995) has undergone important critiques for its deficit-oriented perspective toward families of color, families living in poverty, and families of diverse cultural and linguistic backgrounds, its implications for equity are significant. If the comprehension of complex spoken language is a key element for literacy development, but is not a core element of education, then students who frequently engage with such language outside of school will be more likely to develop that skill. Factors such as family poverty, early vocabulary, and parents' reading patterns have been shown to be influential predictors for below-average oral language abilities in middle school (Law, Rush, King, Westrupp, & Reilly, 2018); language background was not provided in this study.

Instruction that supports oral language development in both EAL and non-EAL learners can support learning outcomes, especially in literacy. Currently, speech-language pathologists and English-as-a-second-language (ESL) instructors are the educational professionals who most commonly support language comprehension in school settings; yet many students – both L1 and L2 users of English – do not have access to such services. With regard to instructional impact, Clarke, Snowling, Truelove, and Hulme (2010) performed a 20-week randomized control trial for 8- and 9-year-old children (language background was not provided) with comprehension-related reading exceptionalities using three different interventions: text comprehension intervention, oral language intervention, and an intervention that combined the two. A follow-up test was administered 11 months after the intervention period, and the group with the largest long-term gains was the oral language intervention group. Bowyer-Crane et al. (2008) found that oral language interventions – specifically in vocabulary and grammar –proved effective in preschool and kindergarten-aged students in areas of receptive and expressive language (participants' language background was not provided). A three-year course of oral language interventions improved French prekindergarten students to improve their ability to detect inconsistencies in a story, make logical inferences, develop a situational (mental) model, and understand story structure (Bianco et al., 2010).

In summary, the broad construct of oral language ability, and its receptive component known as listening or language comprehension, play critical, important roles in successful reading comprehension. The relationship between the oral constructs and reading grows over the developmental trajectory. While not usually a prime focus in mainstream K-12 educational settings, instruction targeting oral language can support the development of reading comprehension skill. The present study examines the relationship between reading comprehension and fine-grained linguistic features of children's oral language; it also compares this relationship to that of reading and writing, discussed next.

2.1.2 The connection between writing and reading

The processes of reading and writing share cognitive and linguistic foundations including pragmatic metaknowledge (understanding the purposes of reading and writing), metacognition (monitoring one's language comprehension and production), domain/content knowledge, knowledge about text attributes (including word, sentence, and text levels), and procedural

strategies for predicting, questioning, and drawing on background knowledge (Fitzgerald & Shanahan, 2000). Shanahan (2016) outlines three frameworks that have been widely applied to research on the relationship between reading and writing. First, the shared cognitive foundation, described immediately above, focuses on the knowledge and skills that are common to both reading and writing. Second is the sociocognitive model, which focuses on reading and writing as communicative acts. For example, it emphasizes that a reader is in conversation with the author, and that reading and writing are social acts occurring within social and political contexts. Third, the combined-use framework conceptualizes reading and writing as practical tools that are used in combination to achieve a goal. These frameworks can also be more finely differentiated, for example, as theories informed by constructivism, transactionalism, socioculturalism, structuralism, and reader response, to name just a few (Hodges, Feng, Kuo, & McTigue, 2016). The strengths of these diverse perspectives can be integrated by focusing on both reading and writing as text-based discourse processes that involve "reading, comprehension, communication, dialogue, argumentation, and of course, language" (McNamara & Allen, 2017, p. 363).

Shanahan (2016) reports that early studies examining the relationship between reading and writing found no more than 50% of shared linguistic variance. These early studies tended to include only one variable each for reading and writing (Fitzgerald & Shanahan, 2000). As studies began to include multiple measures for each construct, estimates of shared variance rose to 65% for text-level constructs and 72-85% for word-level constructs (Berninger et al., 2002). A cross-sectional study examining the cognitive correlates of writing from grades 1-4 (Decker, Roberts, Roberts, Stafford, & Eckert, 2016) found that the most important predictors of writing skill in grades 1 and 2 are low-level perceptual and motor and attention skills (letter-word identification and visual matching), while the strongest correlates in older grades were reasoning (concept formation), verbal language comprehension skills, and working memory (participants' language background was not provided in the study). Decker et al.'s findings suggest early writing skills are focused at the phonemic and letter levels, but later skills draw on complex reasoning and language skills, as well as the ability to mentally hold and process information. In the primary years, the reading – writing relationship is characterized by word-level factors, but as children mature, commonalities across sentence- and discourse-level factors supersede those across word-level reading and writing factors (Shanahan, 2016).

Even before formal literacy instruction begins, children with oral language impairments lag behind typically developing age-peers in writing performance. Puranik and Lonigan (2012) found that preschoolers (language background was not provided) with oral language deficiencies exhibited lower writing scores on name writing, letter writing, and word spelling; students whose language impairment is specific (cognitive scores within normal range) tended to have higher writing scores than students whose impairment is general (cognitive scores below normal range). The authors conclude that nonverbal cognitive abilities are related to writing ability and are moderated by oral language ability.

Ahmed, Wagner, and Lopez (2014) found in a longitudinal study from grades 1 through 4 (language background was not provided) that reading factors tend to influence writing for word- and text levels of language, but a model with bidirectional causality fit the data best at the sentence level. Tong and McBride (2016) corroborated this conclusion in a longitudinal study of Cantonese-using children from ages 4 through 12, finding that writing and syntax skills are bidirectionally related, with each showing evidence of supporting the other, and reading comprehension skill mediates this relationship. Berninger and Abbott (2010) followed two cohorts of English-proficient children in elementary school to determine the relationships between reading, writing, speaking, and listening. Across grades, all four modalities shared substantial variance, although ample variance was unique and unexplained by other factors. They found bidirectional interrelationships (where each contributed unique variance to the other) between listening and reading comprehension and between reading comprehension and written expression in most grades. However, receptive and expressive oral language did not contribute unique variance to one another, suggesting they may draw on a common language core, aligning with the studies cited earlier on the relationship between receptive and expressive language. The authors conclude that the findings support a four-factor theoretical model of language. Shanahan (2016) summarized the relationship between reading and writing over time and found that it tends to show bidirectional causality, with growth in each able to positively influence the other.

Less skilled reading comprehenders have shown to have difficulty producing quality narratives regardless of whether the modality of language production is spoken or written, as shown by Carretti et al. (2016) in a study of L1 Italian-using children ages 8-10. In their study, which utilized Italian language materials, more and less skilled comprehenders were matched on grade, type of school, estimated IQ, and word decoding. Both groups were asked to provide narratives

of two cartoon strips: one verbally and one in writing. Regardless of whether task modality was oral or written, the more skilled comprehenders' narrative output was more syntactically complex and lexically rich than the less skilled group. The more skilled group used causal connecting words (e.g., "because") in their narratives, while the less skilled group tended to use more additive connecting words (e.g., "and"), suggesting that the former understood causal relationships between story events, while the latter group's comprehension was limited to remembering a series of events. The two groups did not differ in their adherence to the task requirements, nor did their narratives differ significantly in length or incorrectly spelled words (for the written narrative). Differences in written narrative output were not related to foundational writing skills, as both groups performed equally well on writing speed and spelling measures.

In terms of demographics, lower SES status has been shown to negatively predict writing quality, after controlling for other demographic variables, but this difference disappeared once vocabulary and literacy skills are accounted for in a samples of L1 Korean-using Grade 1 students (Kim, Puranik, & Otaiba, 2015), highlighting the relationship between vocabulary, literacy skills, and SES. This study also found that low SES also negatively predicts writing productivity, with lower SES students producing approximately 50% fewer ideas after all other variables were accounted for. Such a discrepancy potentially relates to the different forms of background knowledge that students from lower SES backgrounds bring to school. Despite these findings, differences were not found in growth rates between SES groups.

A meta-analysis of studies about instructional interventions targeting the reciprocal relationship between reading and writing (Graham, & Hebert, 2011) found that writing about texts, explicit writing instruction, and increased time-on-task on writing can improve reading comprehension skills. Writing instruction combined with generative writing practice (rather than copy/correct practice) can result in significant gains in reading achievement, after controlling for students' demographics, vocabulary, previous achievement, and the amount of reading and writing instruction, as Coker, Jennings, Farley-Ripple, and MacArthur (2018) demonstrated with a sample of Grade 1 students (9% learning EAL). A meta-analysis of balanced literacy instruction in nine program types (e.g., early literacy, literature-based, strategy instruction) demonstrated significant improvements in both reading and writing, except for reading improvement in content area-balanced literacy programs (Graham et al., 2018). Across the elementary grades, connecting

reading and writing instruction at the skill level is more effective than teaching reading and writing skills that are not related to one another (Abbott, Berninger, & Fayol, 2010).

In the following sections, the specific factors of syntax and lexis, across language domains, are discussed.

2.1.3 Specific factor of syntax

Syntactic ability is the capacity to understand grammatically complex sentences. Comprehension of sentences written with basic and difficult syntax can explain 12% of variance in Grade 5 Danish students' text comprehension (1% of the sample did not use Danish at home), beyond decoding and vocabulary controls (Poulsen & Gravgaard, 2016). For example, less skilled Grade 2 and Grade 4 readers (language background was not provided) have shown difficulty parsing subject and object in complex sentences with relative clauses that have subjects in a near-end position (Stein, Cairns, & Zurif, 1984) such as "The dog bites the cat that the ball hits" (p. 307). In sentences requiring constituent command constraint like "The pig stood near the cow after jumping over the fence" (p. 309), less skilled readers are more likely than skilled readers to think that the cow is the animal that jumped (rather than the pig).

Syntactic awareness, as measured by an oral word-order sentence correction task, has been shown to predict reading comprehension gains in L1 English-using children from Grade 3 to Grade 4, with no mediation by word reading, after controlling for both verbal and nonverbal factors (Deacon & Kieffer, 2018). Syntactic knowledge has also been shown to be a significant predictor of reading comprehension in L1 English-using adolescent students (Brimo et al., 2017). Using French-language materials, Demont and Gombert (1996) longitudinally tested the contributions of phonological and syntactic awareness in predicting reading fluency and comprehension of young French children, while holding IQ and vocabulary constant. They conclude that after Grade 2, only syntactic awareness makes a significant contribution and accounts for substantial variance in comprehension.

Bowey (1986) found that more and less skilled readers in 4th and 5th grades, as defined by word-decoding skill (all were L1 English users), differed significantly in syntactic control; this relationship was stronger than that of syntactic control and reading comprehension. Bowey (1986) argues this may be due to a higher-order language ability, or metalinguistic ability, that

the syntactic and decoding skills share. In a review of literature on the contribution of grammatical knowledge to reading, however, Bowey (1994) argues that research to date has not clearly demonstrated that deficits in grammatical awareness are not simply more "general delays in language development or other skills" (p. 123), and that syntactic measures may to confound grammaticality awareness, semantic processing, and comprehension monitoring. Providing evidence for this argument, Gottardo, Stanovich, and Siegel (1996) found that when holding verbal working memory and phonological sensitivity constant, syntactic processing only accounted for 1.3 to 1.5% of unique variance in reading comprehension of Grade 3 students (all L1 English users) who had no reported language difficulties.

Children aged nine with comprehension difficulties (language background was not provided) who were matched with typical comprehenders on chronological age, decoding ability, and nonverbal ability had significant differences in syntactic awareness (Nation & Snowling, 2000), pointing to a general language processing difficulty that includes semantic and syntactic weakness. The less-skilled comprehenders had more difficulty correcting the word order of sentences, especially those in passive voice or reversible (subject and object both being animate). However, due to the nature of the grammaticality task, they conclude that the less-skilled comprehenders' impaired syntactic awareness may be grounded in general language processing difficulties that encompass both semantic and grammatical difficulties, which aligns with Bowley's (1994) critique. Cain's (2007) study of L1 English-using children ages 8-10 concurred, finding that the variance shared between syntactic awareness (as measured by grammatical correction and word-order correction tasks) and reading comprehension may be attributable to language and memory skill rather than a unique syntactic factor.

In addition to these points about the challenge of defining a unique syntactic construct, deficits in syntactic awareness might not prohibit reading comprehension. Tong, Deacon, and Cain (2013) matched Grade 4 less- and more-skilled reading comprehenders (all from homes where English was the predominant language) on word-reading accuracy and speed, vocabulary, nonverbal cognitive ability, and age. The less-skilled readers performed less well on tasks that just tapped syntactic awareness and morphological awareness, but there was no difference between groups on a task that tapped syntactic and morphological awareness simultaneously.

Studies examining syntactic skills in students learning EAL have found mixed results regarding the predictive power of syntax in reading comprehension, and whether syntactic abilities transfer across languages, but the importance of syntax may grow in adulthood. In a large sample of Grade 6 students who were learning EAL, syntax predicted 7%-19% of variance in a range of reading and spelling skills (Siegel, 2008). Lesaux et al. (2006) found that among Grade 4 skilled reading comprehenders, those who were learning EAL scored lower on syntactic awareness than non-EAL peers; however, despite lacking syntactic proficiency in English, the EAL participants were able to skillfully comprehend what they read. Urdu-English bilingual children in kindergarten have demonstrated significantly greater syntactic awareness, in terms of sentence grammar correction, than monolingual peers, and this capability extended to both languages (Davidson, Raschke, & Pervez, 2010). Yet Gottardo (2002) found that Spanish-English bilingual children's (ages 5 through 8) scores on most syntactic categories did not correlate across languages; her results also suggest that vocabulary and syntax were only correlated within each language. In a sample of adult EAL learners, syntax, and vocabulary predicted 38%-79% of the variance in reading comprehension, with syntax explaining slightly more variance (r=.85) than vocabulary (r=.79) and both variables highly correlated (r=.84) (Shiotsu & Weir, 2007).

One of the strengths of NLP is its capacity to analyze human grammar. For example, Lu (2010) developed a syntactic complexity analyser to examine the complexity of grammar of L2 writing. The units of analysis Lu's program are T-units (minimally terminable unit, e.g., a clause with all subordinate clauses, but not two clauses connected with a coordinate), clauses (a structure that includes a subject and verb), and sentences (which are equivalent to roots). Applications of Lu's grammatical complexity analysis include automated assessment of the spoken language proficiency of L2-English-using children ages 8 and older (Hassanali, Yoon, & Chen, 2015), mapping of children's (ages 5-7) spoken sentences onto a developmental level scale for L1 acquisition (Lu, 2009), and detection of young children's deception or truth-telling through syntactic analysis (Yancheva & Rudzicz, 2013).

Other grammatical features assessable through NLP are the extent to which a sentence is left-branching, that is, having more words and clauses before the main verb (Yngve, 1960). This left-branchingness was included as part of a significant grammatical complexity predictor in the automated scoring of essays written by university freshmen (McNamara, Crossley, & Roscoe, 2013); language background was not provided. On the other hand, placing more words before the

main verb was found to not significantly relate to quality essay writing by L2 English users (Crossley & McNamara, 2014); in this study, the only grammatical complexity aspect that was positive correlated with essay quality was having fewer incidences of all clauses. This is intuitive, though, as having fewer total clauses may be associated with having longer, more grammatically complex clauses. While these studies examine the relationship between grammar and overall quality in student writing, the present study examines the relationship between grammar and reading comprehension.

2.1.4 Specific factor of vocabulary

Vocabulary knowledge refers to the storage and accessibility of lexical representations. The phonological representation of a word is connected to, but distinct from, its semantic representation (Ouellette, 2006). Vocabulary knowledge includes two integrated components: lemma (part of speech and meaning) and lexeme (morphology, spelling, and pronunciation) (Levelt, 1989). These components have been proposed to exist for both L1 and L2 vocabulary storage, but for EAL learners, first the L2 lexeme information and L1 lemma information is established, followed by L2 lemma information, as summarized by Jiang (2000). Vocabulary breadth (the number of lexical entries) can be distinguished from depth, which refers to the extent of learner's semantic representation, or how well each word is known (Ouellette, 2006). Perfetti and colleagues' work on the lexical quality hypothesis (e.g., Perfetti, 2007; Perfetti & Adlof, 2012) suggests lexical depth (including orthography, phonology, morphology, and semantics) rather than breadth (the number of words that are known) is the key factor supporting successful reading (c.f., Gottardo et al., 2018).

In a cross-sectional study of L1 English-using Grade 1 and Grade 6 students by Ouellette and Beers (2010), oral vocabulary predicted reading comprehension only in the older group, when phonological awareness, decoding, irregular word recognition and language comprehension were held constant. Oral vocabulary was also a significant predictor of irregular word reading (i.e., words that cannot be sounded out) for both age groups, and of decoding in Grade 6 (but not Grade 1, where decoding was predicted only by phonological awareness). The authors suggest this is attributable to the lexical restructuring model (Walley, Metsala, & Garlock, 2003) which posits that growth in a child's lexicon, especially in high density lexical neighborhoods (which are clusters of words that differ by only one phoneme) supports phonological awareness

development and thus more efficient decoding. However, Ricketts, Nation, and Bishop (2007) suggest vocabulary uniquely predicts irregular word reading and comprehension, but not decoding accuracy, pseudoword decoding, or reading of individual regular words, in L1 English-using children ages 8-10.

Ouellette (2006) used four vocabulary measures in a study of typically developing, L1 English-using Grade 4 students, allocating expressive and receptive vocabulary tasks to breadth (total size of one's lexicon) and a word-defining and synonym-identification tasks to depth (how well the words in the lexicon are known). Decoding pseudowords was predicted only by receptive vocabulary breadth, again relating to the lexical restructuring model. Expressive vocabulary depth was the strongest predictor of single-word reading, which may be related to word-retrieval functions. Reading comprehension was predicted by all four vocabulary measures, with vocabulary depth explaining an additional 8% unique variance when entered last in a regression equation after age, nonverbal IQ, visual word recognition, decoding, and the other vocabulary measures.

Tannenbaum et al. (2006) examined the contributions of lexical depth, breadth, and fluency (how quickly word meanings can be accessed) toward Grade 3 students' reading comprehension (language background was not provided). Together, the three elements predicted 50% of the variance in reading comprehension, with depth and fluency forming a factor that contributed 19% unique variance, breadth contributing 2% unique variance, and 29% of variance being common to the two factors. Vocabulary depth has also been shown to predict L1 English users' (ages 10-11) ability to make global cohesion inferences but not local cohesion inferences (Cain & Oakhill, 2014), which suggests global inferences rely on the learner's semantic network, which is elicited by vocabulary depth measures.

Tunmer and Chapman, two of the original developers of the Simple View of Reading, investigated contributions of vocabulary to Grade 3 learners' (language background was not provided) reading comprehension development to see if the *language comprehension x decoding* model needed revision (2012). They were specifically concerned about the independence of the decoding and language comprehension components, prompted by studies such as Ouellette and Beers (2010), which suggest that vocabulary development supports the development of decoding skill. When vocabulary was included in the model, it loaded onto the linguistic comprehension

factor, which had a strong positive influence on the decoding factor (which itself consisted of a pseudoword reading and word recognition measures). Thus, they recommend relaxing the assumption of independence between the two components, addresses this question raised in an earlier section on the role of language comprehension in reading comprehension and how decoding relates to both.

In the elementary years, the relationship between vocabulary growth and reading comprehension growth is not completely understood. A study tracking primarily L1 English-using learners from Grade 1 to Grade 4 (Quinn, Wagner, Petscher, & Lopez, 2015) suggests vocabulary and reading comprehension have a unidirectional relationship, with vocabulary growth supporting reading comprehension growth, but not vice versa. Verhoeven, Leeuwe, and Vermeer (2011) found in a longitudinal study of linguistic diverse children in Dutch schools that vocabulary growth and reading comprehension growth were reciprocal in grades 1 and 2, but, like the findings by Quinn et al. (2015), by Grade 3 vocabulary becomes more autonomous as it continues to support reading comprehension development. Vocabulary was quite stable in its growth over time, and, as described for other studies, vocabulary predicted decoding. The longitudinal results suggest that decoding also supports vocabulary development. For both EAL and non-EAL learners in early elementary school (grades 1 and 2), English vocabulary depth was a significant predictor of reading comprehension initial status but did not predict growth rate (Proctor, Silverman, Harring, & Montecillo, 2012). In this study, EAL learners' vocabulary in their L1 (Spanish) did not predict L2 reading comprehension; however, the authors acknowledge that these EAL learners had not ever received formal instruction in Spanish.

A meta-analysis of the impact of vocabulary instruction on reading comprehension (Elleman, Lindo, Morphy, & Compton, 2009) found that vocabulary instructional interventions had three times the positive impact on the reading comprehension skill for students who were less skilled readers than those without reading difficulties. In terms of vocabulary outcomes, comparable levels of growth were demonstrated across reading skill levels. Effective instruction can include intervention based on conversation and interaction in both general oral language and academically specific vocabulary beginning in early childhood and continuing through adolescence and early adulthood (Holahan et al., 2018).

2.1.4.1 Vocabulary richness

Vocabulary depth and breadth, described above, concern vocabulary range, that is, a person's total lexicon. Vocabulary richness, also known as lexical diversity, refers to the diversity of vocabulary produced in written or spoken language. Traditionally, this has been measured in a given sample of language as the number of different words divided by the number of total words, known as the type-token ratio. The type-token ratio value ranges from near zero to 1, with a value close to zero occurring if the same word is repeated consecutively, and a value of 1 occurring where every word in the sample is different. Type-token ratio is typically applied to the analysis of the difficulty of reading materials, but also has been shown to be a contributing factor to children's narrative writing (Cameron et al., 1995). Problematically, though, the standard type-token ratio (across the entire length of a language sample) will always decrease as the length of the narrative increases. Thus, it is not necessarily an adequate metric for vocabulary richness (Chipere, Malvern, Richards, & Duran, 2001). Wood, Bustamante, Schatschneider, and Hart (2019) suggest that the simple count of number of different words in the writing of elementary students (all of whom reported using English at home) has a moderately strong relationship with receptive vocabulary, and that considering length (e.g., type-token ratio) does not improve the strength of association.

The moving average type-token ratio (Covington & McFall, 2010) addresses this problem by using "windows" of words, usually from 10 to 50 words in length. To use a window of 20 words, for example, the type-token ratio is taken for the first 20 words of the narrative, then the window moves one word forward, and the process repeats. The averages of all the windows' type-token ratios is the moving average. Kapantzoglou, Fergadiotis, and Auza Buenavides (2019) found the moving average type-token ratio to be the least biased lexical diversity measure among the four they tested with L1 Spanish-using children with and without developmental language disorders. Two other metrics used to evaluate lexical diversity are Honoré's Index, a function of the number of words used only once, and Brunet's index, a function of the number of different words and narrative length. Brunet's index has lower values for greater lexical diversity, therefore inversely relating to type-token ratio.

2.1.4.2 Subjective and objective ratings of vocabulary

Subjective vocabulary ratings have traditionally been gleaned through surveys, and now they can be gathered through online crowdsourcing (also a form of surveying, but less labor intensive for a research team). Several subjective ratings pertain to the present study and the NLP tools used to analyse participants' productive language. Age of acquisition refers to the age at which an L1 user typically learns a word well enough that it forms a meaningful part of the lexicon; i.e., it is the age at which the word contributes to one's language and memory processes (Stadthagen-Gonzalez & Davis, 2006). Word familiarity measures one's perceived frequency of exposure to a word. Another subjective rating, imageability, refers to the ease with which a mental image of the word can be generated. Word frequency is not a subjective measure, but rather is an objective measure of the number of times a word appears in a corpus of natural language. As to the latter, an example is Brysbaert and New's (2009) frequency corpus, which is based on a corpus of movie subtitles that includes 51 million words. Essentially, these scholars argue subtitles form a more natural corpus than, for example, language extracted from internet chats, which have also been drawn upon for this purpose.

Familiarity, frequency, and imageability are positively correlated, and all correlate negatively with age of acquisition (Gilhooly & Logie, 1980). High frequency words are read and processed more quickly than low-frequency words, for L2 learners as L1 learners, as shown by Kim, Crossley, and Skalicky's (2018) study of adolescent and adult L1-Spanish users as they read English-language texts. As students progress through compulsory schooling, the frequencies of verbs and adjectives in their writing has shown to decrease, while the frequency of nouns increased, as demonstrated by Durrant and Brenchley (2018) using a corpus of writing samples gathered from children ages 6-16 (approximately 13% classified as speaking EAL). Nine-year-old children (language background was not indicated) have shown to read words identified as having older age of acquisition with significantly less accuracy; however, when words were matched on age of acquisition, length, and frequency, and they differed only in imageability, only less skilled readers struggled to read the low imageability words (Coltheart, Laxon, & Keatin, 1988).

2.1.4.3 Word specificity, similarity, and ambiguity

Another set of vocabulary metrics used in NLP relates to the specificity of word use, the ambiguity, or number of senses, in word usage, and the similarity between all words used in a language sample. The WordNet corpus (Miller, 1995) was employed to examine these values in the current study. Specificity of word use in the context of WordNet refers to the "depth" of a given word: how many specificity levels, or nodes, from a root hypernym, is the given word (hyponym)? For example, given the hyponym "nose" (using WordNet's first definition, "the organ of smell and entrance to the respiratory tract; the prominent part of the face of man or other mammals"), the hypernym chain shows a depth of 8 levels from that specific word (the hyponym, "nose") to the root (the hypernym, "entity") that is the most general and least specific: nose—> chemoreceptor —> sense organ —> organ —> body part —> part —> thing —> physical entity —> entity.

This specificity metric is typically used to calculate WordNet's similarity analysis, but its use as a standalone metric for the purposes of the present study is questionable. To illustrate this point, a student's use of "nose", a high-frequency vocabulary word, will have a WordNet specificity value that is higher (more specific) than "chemoreceptor," which has a much lower frequency. Along these lines, Lewis (2002) points out that "cow" has a depth of 13 nodes, while "horse" has a depth of 10 nodes, even though both concepts seem to have the same level of abstraction to the average observer. A weighting function would be useful for such situations (*c.f.,* Wang & First, 2011). Nonetheless, the word specificity metric is included in the present study, as the average path from hyponym (given word) to hypernym (semantic root).

WordNet's similarity analysis examines the relatedness of different concepts (expressed as words, within the same part of speech). Similarity is determined by identifying the least common subsumer (LCS) of two concepts; the specificity of the LCS itself also plays a role in measuring similarity (Pedersen, Patwardhan, & Michelizzi, 2004). For example, "animal" is an LCS of concept A ("kangaroo") and concept B ("koala"), but "marsupial" is a more specific LCS of these two concepts than "animal." The level of specificity of the LCS is known as the information content. WordNet uses several different methods to measure similarity, all of which are based on the basic idea that the more common information the two concepts share, the more similar they are (Meng, Huang, & Gu, 2013).

WordNet offers several methods to analyze similarity. The Resnik method uses the basic information content value when calculating similarity. The Lin method, which scales the LCS information content by the sum of concept A and B's information content. The JCN method subtracts this sum from the information content of the LCS. These methods are combined with different corpora including Brown and SemCor. The WordNet similarity algorithm has been applied to studies of writing cohesion development for university students writing in their L2 (Crossley et al., 2016) as well as NLP-based analysis comparing the sophistication of writing by adolescents and young adults (language background was not identified), as studied by Crossley, Weston, McLain Sullivan, and McNamara (2011); content word overlap was found to diminish over time.

Word ambiguity to the number of different meanings, or senses, a given word has. Words with more than one meaning are homonyms; if they have more than one related (but not identical) meaning, they are polysemous. For example, the verb "run" can refer to the act of moving at a speed faster than a walk (as a horse runs), to pass quickly in a direction (as a rumour runs), to flow (as a river runs), or to extend in a particular direction (as a street runs); indeed, there are more meanings of "run" beyond these. These movement-oriented definitions of the verb "run" can largely be considered polysemous. However, a "run" on the stock market, or in one's stocking, are more homonyms than polysemous.

Adults tend to use more polysemous words than children do. As children mature, their use of polysemous words grows according to a well-defined pattern that is similar for both L1 and L2 language users, through approximately 5 years of age (Casas, Català, Ferrer-i-Cancho, Hernández-Fernández, & Baixeries, 2018). Children first use nouns, and then verbs – and this generalizes both across languages and across L1 and L2 acquisition (Gentner, 1982, 2006); this pattern may exist in part because verbs have higher polysemy than nouns. Adult L2 learners show increased usage of polysemous words as their language proficiency improves overall; however, a paradox exists where lower frequency words typically are less polysemous, and higher frequency words are more polysemous (Crossley, Salsbury, McNamara, 2010), which makes the less feasible the facile prediction of frequency/ambiguity in terms of language development. For example, Crossley et al. (2011) found that university freshman used fewer polysemous words in their writing than Grade 9 students, while the younger group used more

frequently occurring vocabulary than the older group did. The present study includes WordNet's ambiguity metrics by including the number of different meanings for each word.

2.1.4.4 Word sentiment and affect

It is well known that emotions play a pivotal role in literacy, from encouraging memory encoding during reading, to influencing the affect of writing output, to informing cognitive processing. For example, in general, positive emotions encourage creativity and inferential connections, while negative emotions lean toward a local focus and procedural thinking (Bohn-Gettler & Rapp, 2014). Less is known about how the sentiment of the language students produce relates to their literacy achievement. Stanford's sentiment analysis tool (Socher et al., 2013), originally trained on online movie reviews, is a prevalent tool for predicting the valence (ranging from negative to positive) of language and was used in the present study. The advancement this tool provides is that it combines analysis of sentence-level grammar and the sentiment of individual words, thus allowing differentiation between "The movie was terrible," and "The movie was not too terrible." Stanford's sentiment analysis tool has not been applied frequently to children's language, although it has been used as part of model-building to analyze middle school students' social media posts for cyberbullying (Lee, Hu, Chen, Tarn, and Cheng, 2018). This tool may suffer from a lack of generalizability as it is highly dependent on the corpus on which it was trained (Harris, 2018).

A second sentiment analysis approach is the Multi-Perspective Question Answering (MPQA) corpus (Wiebe & Riloff, 2005; Wilson, Wiebe, & Hoffmann, 2005), which also analyzes language at the sentence-level. The goal of MPQA is to differentiate "objective" and "subjective" language and then to identify the polarity of the latter. The MPQA has been included as a component of an automatic scoring model forL2 English-using children's picture narration (Somasundaran, Lee, Chodorow, & Wang, 2015), where the authors found that the inclusion of the MPQA features significantly improved their model above traditional lexical and syntactic analysis. According to Ponari, Norbury, and Vigliocco (2018), emotionally valenced (i.e., from unpleasant to pleasant) words tend to be more abstract than neutral words, and they are processed more quickly by the brain. In addition, for abstract words, those that are valenced have an earlier age of acquisition than neutral abstract words; for concrete words, those that are positively valenced are learned before neutral and negatively valence. However, the

bootstrapping role that emotion may have for the learning of abstract concepts appears to disappear after age 9.

Warriner, Kupterman, and Brysbaert (2013) developed a corpus (through crowdsourcing) of norms for approximately 14,000 English lemmas on valence (from unpleasant to pleasant), arousal (from calm to excited), and dominance (from being controlled to being in control). The ratings in this corpus were almost entirely provided by adults and are parsable by gender, age, and educational differences. For the purposes of the present study, distribution in affective norms across education differences are of interest, and to a lesser extent, age, as age in Warriner et al.'s study was divided into <30 and ≥ 30, a partition that does not directly pertain to this work. Warriner et al.'s study found that all three dimensions had slightly higher (pleasant, excited, in control) ratings from younger than older individuals, while higher education levels were associated with higher valence and arousal ratings, but lower dominance ratings. According to Warriner and Kuperman (2015), the English language, generally speaking, skews toward positivity and calmness.

Warriner et al.'s corpus has been applied to research with children in several areas. These include helping to select stimuli to investigate the characteristics of verbal stimuli that will not exacerbate feelings of anxiety for students with specific learning disorders (Haft, Duong, Ho, Hendrenm & Hoeft, 2019), and training a ML model to learn the acoustic features of words' arousal and valence ratings in a sample children ages 8-11 (Asgari, Kiss, Van Santen, Shafran, & Song, 2014). This corpus has also been employed to examine how the valence, arousal, and concreteness of the verbal context surrounding a given word can influence how efficiently a reader recognizes that word, beyond the impact of these factors in that word itself (Snefjella & Kuperman, 2016).

2.1.5 Summary of language and literacy as interwoven skills

This section described the literature on the interrelationships between the four domains of language: reading, writing, listening and speaking. Two core constructs of language, syntax and lexis, were highlighted, as well as several NLP tools for their analysis. Syntax and lexis have shown to be essential factors in all four language domains. While the majority of research operationalizing syntax and lexis as part of oral language construct have used receptive

measures, the present study investigates their role in productive oral language, as well as productive written language.

The studies cited in this section demonstrate that a well-developed language base can enable successful reading comprehension, when combined with sufficiently developed decoding skills. Previously, decoding, and its antecedent, phonological awareness, were considered to be independent from general oral language abilities. More recent research suggests they may be components of a broader language construct, further raising the educational importance of oral language assessment, instruction, and intervention. However, children have different language and literacy profiles, and students who are learning EAL may present different language-learning needs. Because the present study includes a large proportion of participants who were born abroad and who are multilingual, and because a focus of the study is the generalizability of the NLP-based analysis for a diverse student body, the next section discusses the unique opportunities and challenges for language and literacy instruction and assessment for students learning EAL.

2.2 Assessing language and literacy for students from diverse language and immigration backgrounds

Internationally, approximately 257 million people live outside their country of birth (Economic & Social Affairs, 2017), many of whom bring new languages and cultures to their host country. The convergence of multiple languages and cultures presents immense advantages, and also substantial challenges, for educational systems. Schools must adapt to changing student demographics and provide culturally and linguistically appropriate educational opportunities and highly effective instruction for an increasingly heterogeneous student population. This is especially important considering that, internationally, students whose home environments do not include the language of instruction tend to lag approximately one academic year behind their language majority peers, although differences are significantly reduced once socioeconomic factors are controlled for (Christensen & Stanat, 2007; Marks, 2005; OECD/EU, 2015). In addition, in North American contexts, students from diverse linguistic backgrounds are more likely than monolingual peers to be at risk of dropping out of school (Snow & Biancarosa, 2003).

Appropriate and effective school-based language support programs can mediate these concerns (August & Shanahan, 2006; Baker, 2011; Garcia & Kleifgen, 2010). Even with effective

instruction, though, students whose home languages differ from the language of instruction take an average of four to eight years to learn English and attain grade-level language proficiency (Collier, 1987; Cummins, 1981, 2008). Relative amounts of time to become proficient in English and develop literacy skills in English may differ depending on students' cultural and linguistic backgrounds (1995). Portes and Rumbaut (2001) and Gunderson (2007) found that many factors impact immigrant students' success in school, including multilingual proficiency, level, length, and nature of acculturation, national origin, family economic capital, and school context. Immigrant families' social and human capital, as well as their educational ambition, can support their children's success in school, but social barriers can impede success. These barriers include how the host country reacts and responds to immigrants of different nationalities, institutional racism, and the prevalence of underfunded schools in immigrant communities.

Age upon arrival in the host country can impact the rate at which students become proficient in the language of instruction and reach comparable reading and academic achievement levels as their L1 peers. Those who immigrate later in their school careers and who do not speak the language of instruction at home are more likely to experience an achievement gap than younger students (Cobb-Clark, Sinning, & Stillman, 2012; OECD/EU, 2015). Collier (1987) found a "sweet spot" of students' age upon arrival: those who immigrated to the U.S. between ages 8-11 tended to progress more quickly than students who arrived when they were younger or older. This may relate to several factors. Students who immigrate when they are older may struggle to attain the sophisticated and knowledge and skills required in the secondary school curriculum using a language they are still mastering. As for younger students, the 8-11 year-old group may have had the advantage of receiving some formal schooling in their first language, while students who immigrated at a younger age may not have experienced the same advantage.

Contradicting this idea is a longitudinal study tracking EAL students from kindergarten to Grade 4 (Lesaux et al., 2007) which concludes that while EAL and non-EAL learners exhibit significant differences in prereading and reading skills in kindergarten, differences in skills (including comprehension) are negligible by Grade 4. More recent work by Lesaux and Harris (2013) suggests that while EAL learners close the gap with non-EAL peers in code-based reading skills (e.g., phonological awareness and decoding), reading comprehension tends to develop more slowly. This is likely related to the longer amounts of time EAL students need in order to develop vocabulary and language comprehension skills in English. Although code-based skills

are easier to teach, interventions should not neglect the development of students' overall language proficiency as this is a critical factor for successful reading comprehension (Dickinson, Golinkoff, & Hirsh-Pasek, 2010), as discussed extensively above.

While language support programs are designed to provide essential linguistic scaffolding for students learning the language of instruction, outcomes vary widely for students participating in such programs. Effective programs can support linguistically diverse students' long-term academic growth as well as their acquisition of the language of instruction (Collier, 1992; Cummins, 1992, 2000; Thomas & Collier, 2002). For example, authentic, in-depth, structured, explicit, thematic, text-based vocabulary instruction has been shown to be effective for students learning EAL (Lesaux & Kieffer, 2010). ESL programs are a common approach to support students who are not yet proficient in English, either through pull-out programs or within mainstream classes. Roessingh's (2004) meta-analysis found that successful ESL programs featured collaboration among educator colleagues, supportive administrators, extensive contact time with students, direct and explicit instruction of language objectives, and an advocacy-centered philosophy. Nonetheless, students who enroll in ESL coursework have been shown to demonstrate lower performance outcomes (Gunderson, 2012) and exit high school with significantly less academic content when English proficiency, individual variables, and school-level factors have been accounted for (Callahan, Wilkinson, & Muller, 2010).

Policies and practices for assessing EAL students for K-12 language support programs vary widely worldwide, from holistic approaches (e.g., utilizing observation and interview data) to approaches based on standardized language proficiency assessments (Sinclair & Lau, 2018). A common approach in North American settings is to use a combination of language proficiency measures (either formal or informal) and academic achievement assessments. However, facets such as item type, language of test, students' first language, test-taking environment, and rater can contribute measurement error that inadvertently influences scores and threatens the validity of these assessments (Solano-Flores & Li, 2008). A great deal of further research is needed to ensure the validity of language and literacy assessment measures that assess the growth and academic achievement of students learning EAL for placement, promotion, and graduation purposes. The next section discusses how technological advances can support that work.

2.3 Technology and validity in language and literacy assessment

The infusion of technology into language and literacy research and assessment presents great possibilities for improving both theory and practice, including the potential to make assessment processes faster and more accurate, to provide finer-grained feedback, and even to improve validity. This section further describes NLP and ML, as well as recent language and literacy research that has incorporated these technologies. First, I provide some background about differences between ML and traditional statistics.

2.3.1 What is machine learning?

Early artificial intelligence programs were built on "expert systems" designed to mimic human decision-making through "if-then-else" commands (Deng, 2018). Such systems could not "learn" from data nor could they handle uncertainty—that is, they were deterministic, not probabilistic. Probabilistic analysis in ML goes beyond frequentist interpretations of probability (that is, how many times a flipped coins lands on heads) but instead focuses on what is believed, understood and can be inferred about the subjects in the environment, their future states, and the link between cause and effect (Chater, Tenenbaum, & Yuille, 2006). Probabilistic analyses utilize conditional probability, or the probability of an event given prior data or evidence. This prior evidence is referred to as "belief", which does not refer to human belief *per se;* it refers to prior information that is considered along with new information as probabilities are calculated. Applied to language learning and language processing, this framework considers grammaticality in terms of likelihood. It differs from a categorical notion of grammatical possibility or acceptability (Manning, 2003), where sentences are either grammatically possible or not. Beyond syntax, the cognitive system itself may operate using a probabilistic model of language connecting comprehension and production (Chater & Manning, 2006).

Probabilistic reasoning is a core underpinning of ML. Leo Breiman, one of the developers of the now ubiquitous random forest algorithm (2001a) offers this philosophical summary of ML (what he calls the "algorithmic modeling" approach):

> …[N]ature produces data in a black box whose insides are complex, mysterious, and, at least, partly unknowable. What is observed is a set of x's that go in and a subsequent set of y's that

> come out. The problem is to find an algorithm $f(x)$ such that for future x in a test set, $f(x)$ will be a good predictor of y. (Breiman, 2001b, p. 205)

There are two fundamental differences between traditional statistics – what Breiman terms the "data modeling" approach – and the algorithmic (ML) approach. According to Breiman (2001b), the philosophy of the algorithmic approach assumes that the relationship between input x and output y is complex – nature's "black box" – and quite possibly unknown and even unknowable. Most data modelling approaches, on the other hand, assume a linear relationship between data, or perhaps a higher-level polynomial relationship. But – what if the relationship is circular, or perhaps an even more complex shape? Certainly, all relationships in nature are not equally complex, and some relationships are indeed linear. Breiman is emphasizing that human thinking tends toward oversimplification: that when x changes by a certain value, we are likely to assume that y changes by a certain value, across values of x and y. He argues that researchers cannot assume to know the mechanism that generated the data, and in making this assumption they reduce the probability of building a highly accurate model: "Unfortunately, in prediction, accuracy and simplicity (interpretability) are in conflict" (Breiman, 2001b, p. 206).

For Breiman, with regard to a truly complex relationship, replacing nature's black box with a somewhat opaque ML method such as a neural network (i.e., another black box) is appropriate. This is not to say that researchers should not attempt to interpret an opaque, complex ML algorithm – it simply means that the simplest, easiest-to-interpret model may not accurately represent the data and therefore a simple model may not be the most appropriate. Nonetheless, when a small subset of variables does predict an outcome accurately, there may not be a need for a complex model. A small model (having few variables) that effectively predicts outcomes can avoid overfitting. Overfitting occurs when a complex model appears to be an accurate estimation for the training data, but as it is quite sensitive to the training data, it does not generalize well to test data. On the other hand, if a simple model attempts to approximate complex data, bias may be introduced, meaning that the model's predicted outcomes are not similar to the actual outcomes. Of course, seeking the most parsimonious modeling solution is not new. To make modeling useful for educators and learners, interpretability of predictors and their utility in practice are key. Further, an early predictor is not always necessarily a useful skill for later intervention, and vice versa; developmental stages must be considered. Nonetheless,

interpretability, of both simple and complex models, is possible and desirable, especially when the stakes of the results are high. The related field of Explainable AI has recently emerged as an important voice in determining the who, what, how, and why of ML algorithm interpretability and explainability (e.g., Tomsett, Braines, Harborne, Preece, & Chakraborty, 2018).

Another important way that the algorithmic, or ML approach differs from traditional statistics is that generalizability is a primary concern. This sense of generalizability differs somewhat from that of traditional statistics, where it refers to adequately sampling the population of interest and using appropriate measures and analysis techniques such that findings can be generalizable from the sample to the population at large. In the case of ML, generalizability does pertain to these issues, but is more focused on the results of a specific testing protocol: the ML algorithm, after training on a given dataset, is tested on how well it can make predictions on an unseen "test" dataset. Specifically, the goal is to avoid overfitting, which occurs when a model appears to beautifully predict an outcome from input data, but once novel "test" data are introduced, the model does a poor job of predicting the new datapoints. ML models can be so powerful that they can easily "fit' a training model explaining 99% of the variance in an outcome variable based on the input variables. However, if that trained algorithm can only explain 50% of the variance when applied to the novel test set, this model can be said to be overfit. Therefore, the ML analytic paradigm prioritizes training an algorithm on a set of data and then testing the algorithm on some unseen test data to determine how well it can predict the test set's outcomes from its input variables. A successful ML model has high predictive capability *when tested on unseen data*. This meaning of generalizability is not in conflict with the traditional concept of sample-population generalizability; ideally, the two meanings can be applied for a truly robust generalizability.

As mentioned earlier, two general forms of ML exist: supervised and unsupervised models. In the supervised approach, researchers delineate the variables of interest and use known classifications or scores against which to gauge algorithmic success. The unsupervised approach, on the other hand, clusters, classifies, or orders data without known labels. Of these two approaches, the supervised approach is more commonly used in educational research. The accuracy of supervised ML approaches depends on the quality and quantity of human-scored input on which it learns, thus an iterative and cyclical human-machine process is recommended (Geigle, Zhai, & Ferguson, 2016).

2.3.2 Machine-learning for language assessment

Assessment of productive language that is based on ML holds promise to improve both theory and practice (Chapelle & Chung, 2010). These algorithms "learn" the patterns and distinguishing features of different samples, apply that learning to novel data, and even self-update with new information. The ML approach to assessment already has wide application, ranging from a veterinary patient case analysis (Geigle et al., 2016) to predicting personality traits through social media texts (Lima & De Castro, 2014).

NLP and ML have been applied widely to the automation of essay scoring by large-scale testing organizations such as Educational Testing Service's e-rater (Burstein, Tetreault, & Madnani, 2013). These applications usually involve developing an algorithm that associates an essay's human-assigned score with a set of NLP-derived linguistic features and then testing the algorithm on an unseen batch of essays to determine how effectively the algorithm predicts the human scores from the linguistic features.

Two lesser known but relevant applications of ML-based assessment relate to the current study: the assessment of cognition and the assessment of reading fluency. Fraser, Rudzicz, and Rochon (2013) developed a tool to elicit spoken narratives of older adults and analyze their lexical, syntactic, and acoustic features in order to predict primary progressive aphasia. Using three ML approaches, these authors achieved an accuracy rate of 87%. Roark et al. (2011) developed a ML-based assessment of speech features that accurately assessed mild cognitive impairment, which can be a precursor to dementia. Two pausing variables from a retell task showed significant difference between healthy adults and those with mild cognitive impairment. An algorithm developed by Lehr, Prud'hommeaux, Shafran and Roark (2012) was able to automatically and accurately classify adults with and without mild cognitive impairment using semantic units recalled from a read story. Similarly, Hakkani-Tur, Vergyri, and Tur (2010) compared human- and ML-scored assessment of how many semantic content units a speaker has uttered when recalling a story, and during picture description; they found a high correlation between ML and human ratings.

Prud'hommeaux, Roark, Black and Van Santen (2011) used ML-based speech assessment to classify children diagnosed with autism spectrum disorder, those with developmental language disorders, and typically developing children using speech data gathered via five tasks from the

Autism Diagnostic Observation Schedule: make-believe play, joint interactive play, description of a picture, telling a story from a book, and conversation and reporting. Classification of the three groups was possible using syntactic complexity and surprisal metrics (unexpectedness of a word or part of speech in a given context); however, since 75% of the subjects with an autism disorder also had developmental language delays, it is unsurprising that the algorithm could not differentiate these two groups.

Ample research has investigated ML-based assessment of oral reading fluency, a relatively easy-to-measure construct that correlates highly with overall reading comprehension and is considered a bridge to skilled reading (Pikulski & Chard, 2005; Rasinski et al., 2010). Duchateau (2009) created an automated reading assessment for Dutch children using a phoneme recognizer and finite-state transducer word recognizer. Human-human and human-machine agreements (Kappa values) were very similar. Duchateau et al. also synthesized speech for feedback to help students improve their reading. Proenca et al. (2017) found that rate alone was a strong predictor of manually derived oral reading scores, indicating that human raters are likely very focused on rate. Their best fitting model included reading speeds, the rate of false-starts and repetitions, the rate of all disfluencies, and the difficulty based on pronunciation rules.

Bolaños and colleagues' comprehensive studies of fluency (Bolaños, Cole, Ward, Tindal, Hasbrouck, & Schwanenflugel, 2013; Bolaños, Cole, Ward, Tindal, Schwanenflugel, & Kuhn, 2013) used support vector machine classifiers to analyze words correct per minute, total words, number of word repetitions, insertions, and variance in reading rate, as well as 15 prosodic features such as phrasing, emphasis, and tone (operationalized as changes in pitch and duration). The algorithm was trained to classify read speech into four reading fluency levels, resulting in human-machine correlation (.86) that was even higher than human-human correlation (.80). All lexical features played a significant role in classification, while only some of the prosodic features did: average syllable length, the average number of words between two silence regions, the number of silence regions and the number of filled pauses. Work at Educational Testing Service has utilized these features plus content, vocabulary, and grammar to assess constructed response items using a multiple regression method (Zechner et al., 2014). As with other studies discussed here, speaker-level correlations between machine and human scorers were substantially higher than item- or task-level correlations.

Applied to the reading of individual words, supervised ML methods designed to mimic human raters' processes have also shown higher human-machine correlations than human-human correlations (Black et al., 2009, 2011). In Black et al.'s 2009 study, the first step was to extract features that correlate with the evidence that humans use as they assess read speech. Next, those features were mapped to the human evaluators' judgments. Features that evaluators used were pronunciation correctness, fluency of speech, and speaking rate. Functions were derived from each of these three features such as mean, kurtosis, and range. A combination of lasso and linear regression always selected the mean of the binary acceptable pronunciation (the fraction of words recognized as having acceptable pronunciation), and the upper quartile of the square root of the target word start time. Two other features commonly selected were the maximum square root of time recognized as voiced partial words and the upper quartile of the square root of time recognized as any partial words (negative correlations with overall score). The automatic scoring model's prediction errors were less than the human errors.

In related work, Tepperman, Lee, Narayanan, and Alwan (2011) took an approach to assessing discrete word-reading skills for students learning EAL akin to the construct validation process in assessment design. Their goal was to approximate the "subjective" process teachers use to assess their students. To do so, they delineated three classes of variables: 1) evidence, or observable variables, 2) hidden, or latent variables that are unobservable and represent the learner's cognitive state, and 3) underlying variables such as gender, first language, and item difficulty. An assumption made by these scholars is that there is no single true and correct pronunciation norm by which students' word reading should be judged. Their system was designed to determine if the student's pronunciation of a word is correct by using the child's language's phonological trends as a reference; in other words, their study sought to parse out true errors from differences due to accented English. They trained the algorithm on substitutions, insertions, and deletions common to L1 English speakers, Mexican Spanish accents, and predictable reading errors. Compared with a baseline assessment that does not consider multiple pronunciations, the inclusion of underlying variables and L1-specific pronunciation improved the model's accuracy.

2.3.2.1 Bias in machine learning for language assessment

Bias is a concern with any assessment, whether it uses multiple-choice, or constructed response items. For example, assessment items may exhibit differential item functioning, where students

from different demographic groups are favored or disfavored when overall ability is held constant. Human-scored constructed response items can also suffer from bias because different scorers' results are skewed based on the content, accent, spelling, or other feature of the item response. Not all bias is necessarily consequential: if some items favor one subgroup and others favor another subgroup, bias *may* be negligible across the entirety of the test. In other words, if three items favor one subgroup, and three items favor another subgroup, then the impact on both subgroups *might* be considered negligible, although extensive inquiry is needed to ensure it is negligible across ability levels (Chalmers, Counsell, & Flora, 2016; Oshima, Raju, & Flowers, 1997). Alternatively, where bias across items is collective and additive in favor of one group, bias may be non-negligible (Chalmers, Counsell, & Flora, 2016). However, the negligibility of item bias depends on that item's discrimination value, among other factors; confirming a lack of test-wide bias, when some highly discriminating items indeed demonstrate bias, requires a considerable amount of analysis.

There are several research avenues relating to bias in ML-based assessment that merit investigation. While human raters can tire and be swayed by personal biases, machine-based assessment can also exhibit bias by being trained on biased data or data that lack minority subpopulations (Madnani, Loukina, von Davier, Burstein, & Cahill, 2017). For example, Bennett and Zhang (2016) have raised concerns about basing ML scoring algorithms on scores generated by human raters. In essence, supervised ML algorithms are trained on known groups or a known continuous score, and then evaluated on how well they can classify new data into those groups or ascribe a score to the new data. These known groups are often based on a classification or scoring system that humans created, such as rubrics or performance evaluation.

Bennett and Zhang have argued that relatively little is known about how human raters actually score performances, and that training an algorithm on human scores may inadvertently be "teaching" the ML algorithm to learn and adopt human biases. These authors conclude that supervised ML-based automated scoring tools should be critically examined. They give an example of a ML competition held in 2012 which invited developers to create algorithms that could grade essays, with a prize awarded to the developer who could best correlate with human scores. One entrant, whose contribution had a kappa of only .02 less than the winning team's, commented that he had learned from his analysis that the two human scorers did not seem to value the use of sophisticated vocabulary in essay writing. Thus, he removed that feature from

his algorithm, and focused the analysis on the raw number of commas, since that was a strong predictor. The use of commas itself is not a meaningful indicator of quality writing.

The "perfect storm" of concerns related to this scenario are described by Bennett and Zhang (2016) as this: ML is a highly complex technology which is constantly evolving to be even more sophisticated; many ML algorithms are considered proprietary trade secrets and are not disclosed; the algorithms are designed to emulate human-assigned scores, which may not themselves be valid, as the process of human scoring remains somewhat opaque; and, lastly, interrater reliability – with potentially biased human-generated scores – is often provided as the sole validity evidence. Thus, a tension exists between the focus on validity in educational assessment (Messick, 1989; Mislevy & Haertel, 2006) – which methodically connects the educational domain to the assessed construct(s) to the scoring model to score interpretation, and asks if the inferences and actions based on test scores are appropriate – and ML, which can find factors that correlate with "valid" human-designated scores, yet may not include all factors that human experts deem important.

Some ML assessment research has attempted to address this concern, for example Black, Tepperman, and colleagues' work on Project Tball (Technology Based Assessment of Language and Literacy) (Black et al., 2009, 2011). They specifically designed their algorithm to emulate teachers' logic by considering conditional dependencies among the student's language background and potential for accented speech, item difficulty, prior performance, and evidence features that intuitively make sense for assessing word-reading (namely rate, substitutions, insertions, and deletions). Their work also described the human rating process in depth. However, their work focused on assessing single word reading, which may be simpler than assessing oral fluency when reading multiple sentences or an entire passage. Nonetheless, making the process of algorithm development intuitive, and connecting the variables used with educational theory during front-end work, can support validation of ML-based assessment.

Can ML-algorithms find or "understand" differences that are not readily apparent to humans? This would relate more to unsupervised ML, where the algorithm is not given human-designed guidance to identifying differences in the data, but rather seeks out correlations and patterns in the data, or identifies underlying organizing elements in the data, as with factor analysis. This study addresses these related questions by using both supervised and unsupervised ML to

analyse large linguistic feature sets, with the goal of determining if there is a hitherto unreported relationship between fine-grained lexical and syntactic features and reading comprehension. To reiterate the study's research questions:

RQ1. How much variance in reading comprehension can be modelled as a function of productive lexical and syntactic features extracted through NLP? How does the variance explained compare to that explained by models that use traditional lexis and syntax measures? What are the top lexical and syntactic feature predictors for each of the four models (oral/text elicitation by more/less skilled readers)? How do the top predictors differ across these models?

RQ2. Is there significant interaction between the top lexical and syntactic features and students' length of residence in Canada, or their multilingual proficiency in languages other than English, in predicting reading comprehension? If interactions exist, how do they differ across the four models?

RQ3. What latent factors can be identified in the NLP-derived data using an unsupervised ML method? Do the factors differ across the four models? Do the resulting factors predict reading comprehension? How do these predictive relationships differ from those in the first research question?

3 Methodology

The present work is part of a larger study examining children's literacy development through technology-infused self-regulation intervention and assessment (Jang et al., 2018).

3.1 Participants

Students in grades 4-6 (N=172, 48% female) were recruited from ten classrooms in three elementary schools in a metropolitan area of Ontario, Canada. Thirty-nine percent were in Grade 4, 28% in Grade 5, and 32% in Grade 6. Forty percent of participants were born outside Canada, and Figure 2 provides a map of participants' countries of origin with different shades representing the number of participants from each country. If participants immigrated from abroad, they were asked how long they have lived in Canada. All participants were asked to

indicate languages they know (aside from English) and to rate their proficiency in each of those languages from 1-5. Figure 3 represents the relationship between participants' years in Canada, the sum of their reported language proficiencies and whether or not they were born in Canada. A visual inspection of Figure 3 suggests participants who were born in Canada reported a range of multilingual proficiencies, while most participants born abroad tended to report summed language proficiencies above 4. The participant group was highly diverse in terms of their language use. Table 1 indicates the proportion of students reporting specific additional languages. The columns indicate the proportion of the total sample that identified speaking each language as a first language, second language, etc. For example, in the first row, about 24% of the students in the sample reported speaking French as their first additional language, about 15% reported speaking French as their second additional language, and so forth. The table does not sum to 100% because participants varied in how many additional languages they spoke and not every student in the sample indicated speaking a language other than English. Collectively, these figures and table demonstrate the sample's diversity in terms of immigration and language background.

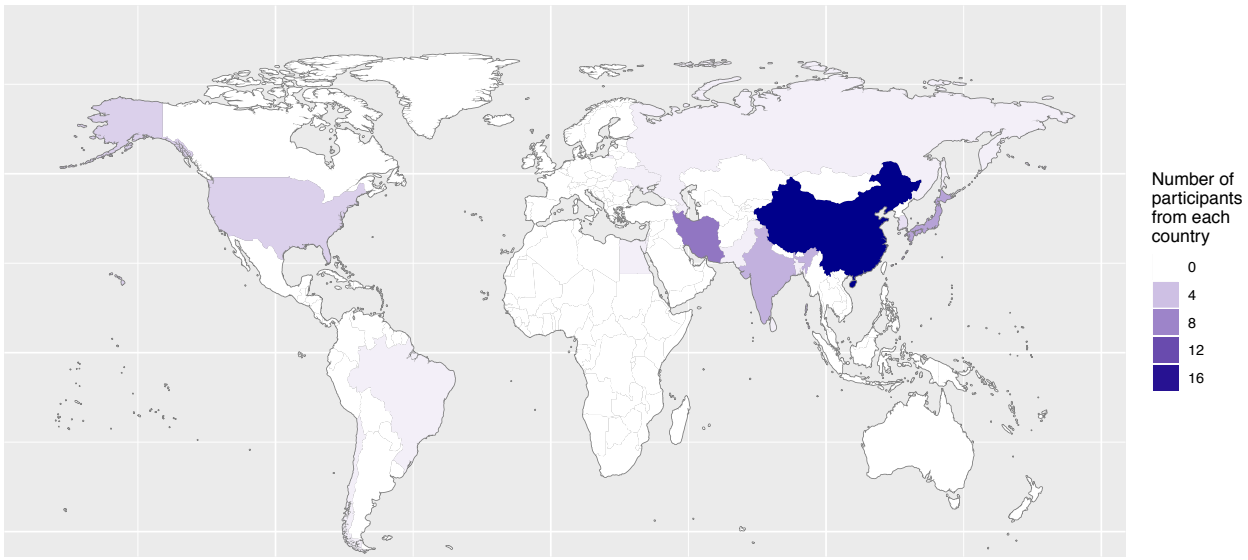

Figure 2. Global map representing countries of birth of participants who were born abroad.

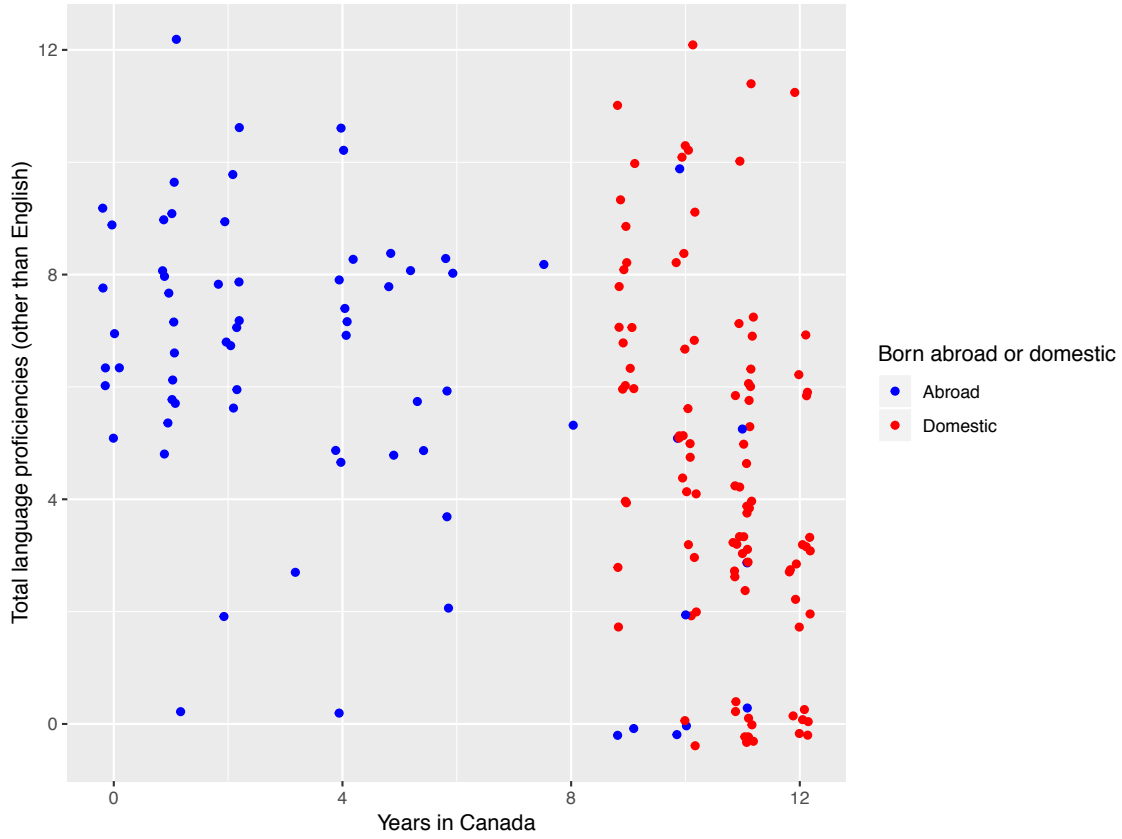

Figure 3. Scatterplot of participants' summed language proficiencies (other than English), their years in Canada, and whether they were born in Canada or abroad

Table 1

Proportion of study participants reporting specific additional languages they know (other than English)

	1st addl. language	2nd addl. language	3rd addl. language	4th - 6th addl. language
French	24.02	14.71	2.94	2.94
Chinese*	20.58	6.86	2.94	1.47
Korean	6.86	.98	.98	
Japanese	4.90	.49		.49
Russian	3.43	.49		
Farsi	2.94			.49
Arabic	1.96		.49	
Spanish	1.47	.98	1.47	.49
Hindi	.98	.98		
Iranian	.98			
Marathi	.98			
Persian	.98			
Portuguese	.98			
Bangla	.49			
Bulgarian	.49			
Filipino	.49			

Greek	.49			.49
Hindi	.49			.49
Italian	.98		.49	
Malayalam	.49			
Marathi	.49			
Polish	.49			
Sign language	.49			
Sindhi	.49			
Sinhalese	.49			
Tagalog	.49			
Tamil	.49			.49
Urdu	.49	.49		
Vietnamese	.49	.98		
German		.98		
Ukrainian		.98		
Dutch			.49	
Gujrati			.49	
Kurdish			.49	
Hebrew				.49

* Chinese also includes Mandarin, Cantonese, and Shanghainese.

3.2 Measures and procedures

3.2.1 Reading comprehension measure

The reading comprehension outcome measure was the Balance Literacy Assessment (BALA), developed by Dr. Jang's research team in both online and paper-based formats to meet teachers' preferences. BALA elicits explicit, inferential, and discourse reading comprehension skills through one narrative and one non-fiction passage. Two versions of BALA were developed: regular and modified. Participating teachers (N=10) indicated which participants would benefit from a modified version of BALA, including students who were currently enrolled in an elementary ESL program, and students who were experiencing substantial difficulty in reading. These participants, as well as all participants in Grade 4, completed BALA- modified. Students in grades 5 and 6 who were not identified as benefitting from the modified version completed the regular version of BALA. Regular BALA consists of 18 multiple-choice items (n=132, α=.80) and modified BALA consists of 17 multiple-choice items (n=109, α=.85). These n-values and reliability figures represent the entire data sample for the two BALA assessments that had been gathered to date. These n-values summed are slightly greater than the total sample in the present study (N=172) due to time constraints, transience, and a lack of consent for some students to participate in the online aspects of the research study. Appendix B provides item-level classical test theory statistics for the regular and modified versions of BALA. Table 2 indicates the complexity of each text.

Table 2

Description of texts used to assess reading comprehension

Genre	Regular version	Modified version
Narrative	"Marilyn Bell" (299 words) Flesch-Kincaid: 6.76 TextEvaluator: Within Grade 7	"Big Sister" (250 words) Flesch-Kincaid: 2.15 TextEvaluator: Within Grade 2
Non-fiction	"Bats" (203 words) Flesch-Kincaid: 6.47 TextEvaluator: Below Grade 5	"Owls" (233 words) Flesch-Kincaid: 3.82 TextEvaluator: Within Grade 2

3.2.2 Text elicitation measures

Participants' written language was elicited through two measures. First, the BALA reading comprehension measure also asks students to (1) make open-ended predictions about what they will read, (2) describe their interest in the text, (3) generate three questions about what they read, and (4) answer a high-level comprehension question. Second, a separate writing task consisted of a video stimulus about children's use of social media followed by the writing prompt, "Is social media mostly good or bad for young people?" Text files of participants' written responses to the open-ended questions about the BALA reading passages and their written response to the social media writing prompt comprise the raw data for the text-elicited linguistic feature set.

3.2.3 Speech elicitation measures

Speech was elicited by Talk2Me, Jr. (Jang, 2019), a web-based tool adapted from Talk2Me, a platform designed to assess language and cognition of older adults who may be at risk for dementia (Komeili et al., 2019). Talk2Me, Jr. uses Talk2Me's web and database architecture, using novel tasks designed to assess elementary students' language, literacy, and cognitive skills. Talk2Me, Jr.'s tasks are designed to be completed with minimal support. Child-friendly visual design includes large font size and uncluttered formatting. To ensure instructions are clearly understood, verbatim audio recordings of each task's instructions (created using Amazon Web Service's "Polly" text-to-speech tool) automatically play at the start of each task. Audio recordings of instructions, prompts, and stimuli can be repeated as many times as participants like.

For tasks eliciting speech, participants click buttons on the screen in order to listen to prompts and stimuli, to begin recording spoken responses, and to terminate recordings (Figure 4 is an example). Oral, written, and clicked responses are stored on a secure server for feature extraction and analysis, and the research team documented any administrative irregularities. For the current study, Talk2Me, Jr. was administered on microphone-equipped laptops in quiet locations of participating schools, and participants wore headphones for audio clarity. Talk2Me, Jr. includes seven tasks, two of which are included in the present study: picture description and story retell. Students' spoken responses to these tasks were transcribed by the research team. The resulting text files comprise the raw data for the speech-elicited linguistic feature set.

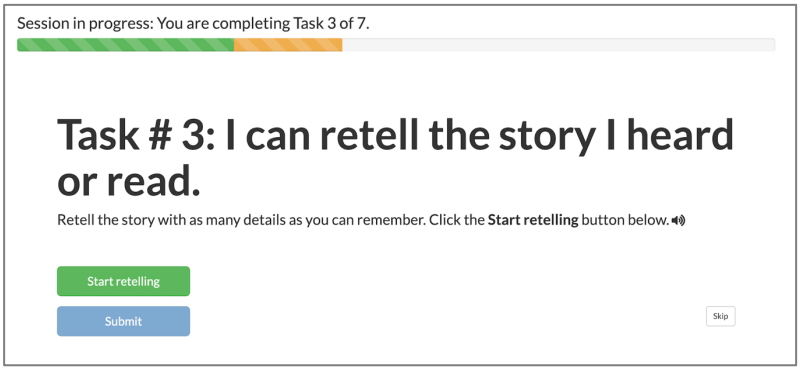

Figure 4. Sample of Talk2Me, Jr. assessment platform interface

3.2.3.1 Picture description

Participants were asked to describe two pictures using full sentences. The full-color pictures are presented in sequence and do not relate to one another. The first picture is a kitchen scene, very loosely modeled on the Boston Cookie Theft image (Goodglass, Kaplan, & Weintraub, 1983). The picture description task used in the present study features a woman preparing food, a boy playing with a dog, and a girl who is about to fall while reaching for cupcakes on top of a refrigerator. In the second image, a police officer falls off her bicycle as a dog chases a cat into her path. For this, and the following task, participants are prompted to click to start and end the recording of their speech.

3.2.3.2 Story retell

Participants were asked to retell two fictional stories with as many details as possible. The first is about a girl arriving late to the first day of school. It is presented aurally (a member of the research team recorded herself reading the story aloud) and consists of 182 words. TextEvaluator, an online tool to evaluate text complexity (Napolitano, Sheehan, &

Mundkowsky, 2015) gauges its complexity within Grade 6, and its Flesch-Kincaid (Kincaid, Fishburne, Rogers, & Chissom, 1975) readability score is 5.15. The second passage is read quietly/silently by participants, who then retell it. This passage consists of 294 words and centers on a playground conflict that is resolved by inviting the antagonist to race. It has a Grade 6 TextEvaluator level but a Flesch-Kincaid score of only 4.69. (This discrepancy may be due to Flesch-Kincaid's basis of syllables, words, and sentences, while TextEvaluator considers many variables such as connections across ideas and text organization.)

3.2.3.3 Demographic self-report.

Participants also completed the Balanced Literacy Learning Inventory, a self-report of demographics and dispositions toward literacy learning. Of these, only two demographic variables are included in the present study. These are whether participants were born within or outside of Canada (if the latter, their age upon arrival), and the languages other than English the participant knows (and self-assessment of their proficiency in each language).

3.3 Data pre-processing

3.3.1 Natural language processing

The raw text files of the text-elicited and transcribed speech-elicited responses were processed using the Core Variable Feature Extraction Feature Extractor (COVFEFE) NLP package (Komeili et al., 2019). COVFEFE offers a multitude of feature extraction pipelines, each of which essentially compiles several linguistic feature extraction programs that otherwise would need to be processed individually. The current study uses COVFEFE's lexicosyntactic pipeline, which, using several corpora and processes outlined in Table 3, extracts metrics related to lexis and syntax. Although this pipeline extracts semantic features (i.e. the meaning and content of language, for example, the cosine distance of meaning between utterances), in the present study such features are not relevant to the research questions nor are they potentially generalizable beyond this study, and they were thus removed. After this processing, 260 features were retained in each the text- and speech-elicited sets.

Table 3

Features extracted through COVFEFE's lexicosyntactic pipeline

Name	Description of features	Citations

Vocabulary richness and range	Corpus of age of acquisition, imageability, and familiarity metrics for all words, as well as nouns and verbs specifically. Type-token-ratio functions include moving average and Brunet's and Honoré's indices. Syllabic counts include per word and Flesch-Kincaid.	Brysbaert & New (2009); Covington & McFall (2010); Chinaei et al. (2017)
Syntactic complexity	Normalized count of clauses, clauses per sentence and per T-unit, complex nominals, T-units, coordinate phrases, dependent clauses, sentences, and verbs; ratios among these. The Stanford Parser extracts counts of 110 grammatical constituents (sequences of parts of speech) normalized by length of sample	Lu (2010); Manning et al. (2014)
Vocabulary specificity, similarity, and ambiguity	Wordnet corpus measures of word specificity: paths from the synonym set of a given word to the root hypernym (broadest umbrella term for given word). Functions include averages and standard deviations of the longest and shortest paths. Ambiguity metrics include mean, standard deviation, kurtosis and skewness. Similarity metrics include those described in section 2.1.6.3	Miller (1995)
Vocabulary sentiment and affect	Warriner's corpus of ~14,000 English words rated on valence (unpleasant to pleasant), arousal (calm to exciting), and dominance (controlled to controlling). Stanford Sentiment Analyzer and MPQA Lexicon corpora measure words affect from negative to positive.	Warriner et al. (2013); Wilson, Wiebe, & Hoffmann (2005); Manning et al. (2014)

3.3.2 Data cleaning

The COVFEFE lexicosyntactic pipeline output results for the 260 lexical and syntactic variables as one comma-separated-value file per input file (that is, one data file for each participant's individual oral or written language sample). These were appended using Python and appended by modality, resulting in a productive oral language dataset containing Picture Description 1 (PD1; n=168), Picture Description 2 (PD2; n=152), Story Retell 1 (SR1; n=151), and Story Retell 2 (SR2; n=140), and a written productive language dataset containing the open-ended reading response (n=154) features, and writing (n=155) features. Concerning the open reading responses, a separate NLP output was produced for each response to the open reading items (participants' reading prediction, description of their interest in the topic, questions they generated, and their response to a high-level comprehension question). Once appended, there were 1567 open reading responses (each with the same 260 lexicosyntactic variables) for 154 participants. It was necessary to reduce the dimensionality of these data and address data sparseness (as not all children completed every single reading response item), since ML does not typically manage missing data well.

To do so, the mean of each of the features of the open reading responses was calculated, per person, with the same feature across all open reading items and across both reading passages. In turn, each feature from this averaged feature set was averaged with the same feature extracted from the writing task output. Missing values were ignored; thus, missing an element for one task

does not negatively impact the score for that domain. For example, the count of all nouns (normalized by the total words in the text file) is one feature extracted from the COVFEFE pipeline. This datapoint was averaged for each participant across all the open reading responses, and then the resulting datapoint was averaged with that participant's noun count from his/her writing response, resulting in a single noun count datapoint that includes information from all the participant's written responses. The same procedure was utilized to average features across the speech-elicited feature sets (two picture description tasks and two story retell tasks). This process reduced the high dimensionality of the datasets and addressed data sparseness by allowing inclusion of participants who may be missing one or more tasks. In addition, it allows interpretation of results at the level of domain (oral and written elicitation) rather than at the item or task level. Because most of COVFEFE's count variables are normalized by length during initial NLP, the difference in length of responses (e.g., for participants who are missing responses to the longer items versus those who are missing responses to shorter items) was not hypothesized to be problematic. After processing, there were two feature sets, each with 260 variables: the oral-extracted set (n=165), and the text-extracted set (n=166). After this process, there were no missing values because all participants completed at least one speech task and one written task. Variables with no variance were removed using the "nearZeroVar" command in R.

The reading comprehension outcome (BALA) was merged with the oral and written datasets, matched by participant ID. The datasets were then divided into the regular and modified version of BALA for analysis (their relatively normal distributions can be seen in figures 5 and 6), since these represent different populations (younger/less skilled readers, and older/more skilled readers), resulting in four datasets, as outlined in Table 4. To reiterate, teachers indicated which of their students would benefit from a modified version of BALA, including students who were currently enrolled in an ESL program, and those who were experiencing substantial difficulty in reading. These participants, as well as all Grade 4 participants, completed BALA-modified. Students in grades 5 and 6 who were not identified as benefitting from the modified version completed BALA-regular. Two participants had scores on the outcome measure were more than 3 standard deviations lower than the mean; they were omitted from the study. Finally, the two self-reported variables from the learning inventory were merged in to the datasets for the second research question: number of years in Canada (age minus the age upon arrival) and total proficiencies in languages other than English, described above.

Table 4

Sample sizes and mean and standard deviations of BALA reading comprehension score (proportion correct) for each of the four datasets

	Version of reading comprehension outcome measure	
	Regular	Modified
Oral-elicited dataset (from picture description and story retell tasks)	*n*=95 *M*=.74 *SD*=.12	*n*=70 *M*=.81 *SD*=.13
Text-elicited dataset (from open reading response and writing tasks)	*n*=99 *M*=.74 *SD*=.12	*n*=67 *M*=.82 *SD*=.13

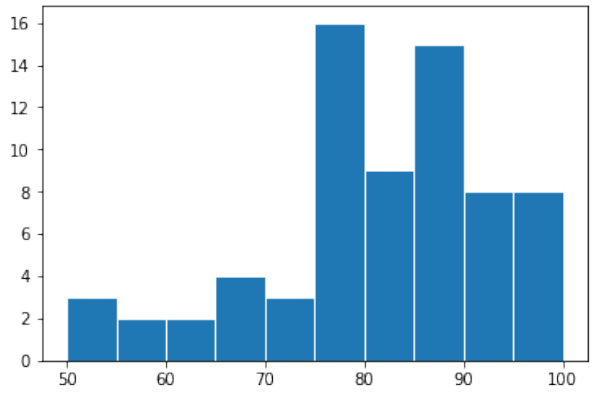

Figure 5. Distribution of BALA reading comprehension scores for modified version (*n*=70)

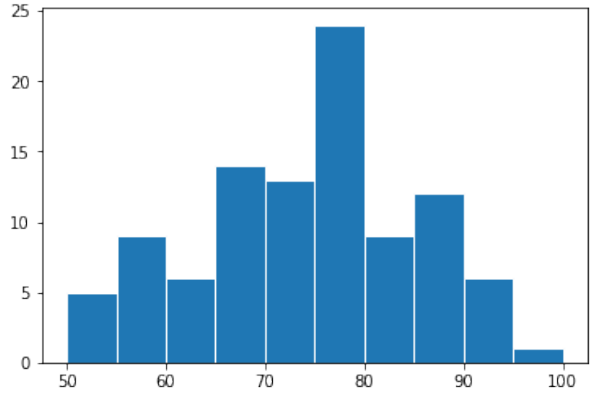

Figure 6. Distribution of BALA reading comprehension scores for regular version (*n*=99)

3.4 RQ1: Supervised ML models with individual NLP features predicting reading comprehension

The first research question uses supervised ML to examine the variance in reading comprehension that can be explained through NLP-extracted lexical and syntactic speech features. For each dataset (oral-elicited/regular, oral-elicited/modified, text-elicited/regular, and text-elicited/modified), eight different supervised ML models were trained, tuned, and tested against a mean baseline. The top predictors were identified through permutation and analyzed and compared across the four datasets. These processes are described in detail below.

3.4.1 ML model-building and testing

The ML analyses for RQ1 are all based on regression, in that they predict a continuous outcome (the reading comprehension score). All were performed in Python (version 3.6.7) using the Scikit-Learn package (version 20.3, Pedregosa et al., 2011). The results of eight ML models, which are described in Table 5, were compared to a mean prediction baseline. The model types (decision tree ensemble, nearest neighbor, support vector machine, neural networks, and a regularized linear method) are described more fully in the following section.

Table 5

Description of eight ML regression models used for RQ1

Model	Description
Decision tree ensemble methods	
Decision Tree	Non-parametric method that predicts an outcome by compiling results across a specified number of independent decision trees
Gradient Boosting	Learns from decision trees in sequence, where the results of initial trees "boost" subsequent results, resulting in higher accuracy
Random Forest	Compiles results from independent decision trees, but only a fraction of k predictor variables (usually $\sqrt{k}$) are used in each tree
Nearest neighbors method	
K-Nearest Neighbors	Predicts the outcome of an unknown datapoint based on the mean of the nearest neighbors in the training data feature space.
Support vector method	
Support Vector	Identifies a linear or non-linear hyperplane in the feature space that captures and predicts the outcome, focusing on datapoints on/outside a margin of a specified distance (epsilon) tracing this hyperplane
Linear Support Vector	Like support vector, but faster as it only implements linear kernels

Neural network method	
Multi-layer perceptron	Learns linear or non-linear patterns by activating a series of hidden layers (neurons) that sequentially transform the input features through a weighting function; at the output layer the loss (error) is gauged and used to revise subsequent learning
Regularized linear method	
Least absolute shrinkage & selection operator (Lasso)	Shrinks unimportant coefficients to zero, through regularization parameter λ which penalizes regression coefficients; greater values of λ shrink a greater number of coefficients to zero

I performed an iterative process of cross-validation for model-building. First, each of the eight models was run using the default parameters. Then, the GridSearchCV method was used to tune and test different combinations of model parameters to find the best-fitting model. For both default and tuned models, the reported results are produced through cross-validation. For this, Scikit-Learn's ShuffleSplit was employed to create 200 random splits in the input data for training (80% of each dataset) and testing (20% of each dataset). In each iteration, the model is trained on the training set, in that it "learns" the relationship between the lexical and syntactic features (input variables) and the continuous reading comprehension outcome. From the patterns that are learned in training, parameters are estimated that best fit the data: these parameters become the algorithm (using an analogy to linear regression, these are akin to regression coefficients – or they indeed may be regression coefficients if a linear model has been specified). Then, this algorithm is applied to the 20% of the dataset that had been set aside for testing. The algorithm processes the input variables of the test set and calculates the predicted outcome based on the parameters that were estimated using the training set. The predicted outcome is compared to the "ground truth", or actual outcome, and the mean absolute error (MAE) is calculated between the predicted and actual datapoints. This process continues on each of the training set/test set combinations, and the MAE across the 200 iterations itself is averaged. This is the metric by which model fit is judged.

Because the four datasets tested here are wide (there are three to four times more variables than individuals in each dataset), the cross-validation method is preferable to reporting a single test-train instance. In small, wide datasets, the nature of the split – that is, which observations are in the training versus testing datasets – can strongly influence the results. To some extent, cross-validation mediates this problem and supports generalizability.

The mean baseline algorithm was the "mean" strategy in "DummyRegressor" from sci-kit learn. This baseline was performed using the same cross-validation process as the other models: 200 different training (80%) and test (20%) sets are created at random, the model is trained on the training data and tested on the test data for each of the 200 sets, and then the mean absolute errors (predicted-expected) are averaged across the 200 sets. DummyRegressor sets the training algorithm to predict the mean value of the dependent variable from the set of independent features. For the test set (within each cross-validation), the algorithm applies what it learned from being trained to predict the mean and attempts to predict the test outcomes. Just as the other models, the absolute difference between predicted values and actual values is summed and divided by the number of observations.

The subsequent section describes the eight different modeling approaches for RQ1 in greater depth.

3.4.1.1 Decision tree ensemble methods

As outlined in Table 5, three decision tree ensemble methods were employed to address RQ1: a basic decision tree ensemble method, gradient boosting, and random forest. "Decision tree" refers to the use of splitting rules (decisions) to segment the predictor space (in this case, the lexical and syntactic linguistic features), *X1, X2, ...XP* into *J* distinct, non- overlapping regions, *R1*, *R2*, *...RJ*. In a regression (continuous) problem like RQ1, these regions are determined by the mean of the response variable (James, Witten, Hastie, & Tibshirani, 2013), which in this case is the reading comprehension score. The properties of each region are determined by recursive binary splitting, or a top-down, greedy approach that minimizes the residual sum of squares (RSS). Top-down, greedy approaches are called such because the best split (i.e., the split that reduces the RSS the most) is selected, without regard for the potential impact on future steps. Unlike classic regression, decision trees do not rely on a linear assumption. The formula applied to decision tree splitting is (from James et al., 2013):

$$\sum_{i:\, x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\, x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$$

where $\hat{y}_R$ is the mean response for the dependent variable in the region R_J based on cut score s in region J. This process of splitting the regions continues with the resulting regions until a stopping criterion is reached. This is demonstrated visually in Figure 7.

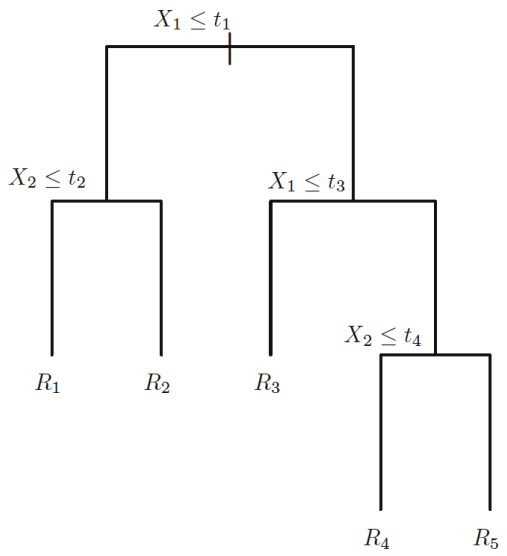

Figure 7. Depiction of a decision tree splitting of *X1* and *X2* creating internal nodes that form R_J regions, from James et al. (2013).

When this process is used on training data, it may result in overfitting. That is, the finely-noded trees make exceptionally good predictions on the training data, but with such a high level of complexity, the model may not predict the outcomes well for the novel test data. Here, pruning is introduced. First, a highly complex ("large") tree is fit on the training data, but then it is pruned back using weakest link pruning. During cross-validation, tuning parameter α negotiates the trade-off between tree complexity and the fit to the training data and is used to minimize the error on test data. Classification trees use the same procedure as regression trees, except the most commonly occurring class is used instead of mean continuous response.

Single decision trees do not have high levels of accuracy and can result in overfitting, but techniques that use an ensemble of trees can improve accuracy (James et al., 2013). Ensemble techniques can take two forms: boosting or bagging. Boosting, as in gradient boosting, is a sequential method where the algorithm learns from its own results at each iteration – that is, errors and the associated residuals of one iteration become the focus of the next iteration.

Bagging, on the other hand, involves bootstrapping a given training sample by taking *B* samples from the training sample, creating large trees for each bootstrapped sample, and averaging the decision tree results.

Random forest improves bagging models by decorrelating the ensemble of decision trees (Breiman, 2001a). Normal bagging always begins with the strongest predictor, that is, the predictor that will reduce RSS the most (James et al., 2013). All trees thus emerge highly correlated, and taking the mean of this multitude of trees will actually be similar to the single-tree model – variance will not be reduced that much. In contrast, random forest iteratively randomizes the predictors that are included in the model. A random subset *m* of total predictors *p* are selected at each iteration, often using $m \approx \sqrt{p}$. Thus, the trees are less correlated, so taking the average reduces more variance than in bagging, and reliability increases.

3.4.1.2 Nearest neighbor methods

Nearest neighbors is a non-parametric ML method that can be used for unsupervised, supervised classification, and supervised regression problems. Generally speaking, nearest neighbors algorithms cluster observations based on their closeness in multidimensional Euclidean space. The *k* of *k*-nearest neighbors refers to the number of neighboring observations to consider around a given query point. Nearest neighbor algorithms are not generalizable but are appropriate for specific problems such as recognizing handwriting. With regard to RQ1, which calls for a supervised regression method, the nearest neighbor algorithm "remembers" the location of each observation in the training data in Euclidean space of predictors *X1, X2, ...XP* (in this case the lexical and syntactic features), as well as the outcome variable (reading comprehension score) associated with each. When the novel test data is presented, the algorithm searches for *k* observations closest to each new observation's predictors and estimates the outcome based on those *k* observations' (in the training data) mean outcome. The error is ascertained by finding the difference between the estimated and actual outcomes.

3.4.1.3 Support vector methods

Support vector machines are a popular and flexible non-parametric ML algorithm that can predict continuous, binary, or multinomial outcomes when the relationship between the predictors and outcome is linear, radial, polynomial, or sigmoid in shape. A defining feature of support vector machines is a hyperplane: a flat subspace that uses one less dimension than the

predictor dimension (i.e., the dimension of predictor variables, in this case the lexical and syntactic features) (James et al., 2013). In the classification problem, a hyperplane separates the predictor dimension into as many regions as there are classification outcomes. With regression problems, the hyperplane helps to predict the continuous outcome variable (reading comprehension score). Here, a predetermined margin of error (the "epsilon tube") is set by the researcher (and can be tuned later). Errors for predictions occurring within these margins are not penalized, while observations that sit outside the margin are the "support vectors" that define the regression function. Tuning the epsilon tube to allow for a greater margin of error reduces overfitting but may also increase bias.

3.4.1.4 Neural networks

A neural network comprises a set of neurons that are inputs and outputs, in between which are one or more hidden layers of neurons. Input values are passed to the first hidden layer. Depending on the value of each input, the importance of each input (its weight), and the bias, or activation threshold, of each hidden layer, the neurons activate to pass information through each hidden layer and eventually the output. Neural network algorithms automatically optimize the weights and biases of the input and hidden layers to minimize error. An illustrated depiction of a neural network is provided in Figure 8.

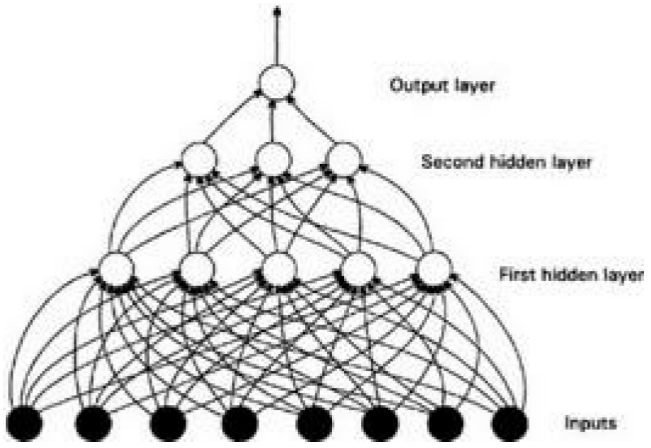

Figure 8. Depiction of a neural network with two layers, from Gurney (2014)

A common method for optimization of neural networks is gradient descent, which uses an iterative, step-wise process to identify the weights and biases that minimize the cost, or error between the actual output and the output that was estimated through the neural network's hidden

layers. Two passes occur in the neural network. First, the forward pass activates the input nodes, passing through the inner, hidden layers, finally producing a set of outputs. The weights, which can be considered a coefficient that represents influence, are fixed during the forward pass. Then, the back pass propagates a signal through the layers in order to adjust the weights and biases, minimizing error by estimating how changes to weights and biases would impact the cost function and making the appropriate changes for the next forward pass (Dao & Vemuri, 2002). Bidirectional models are devised such that each point or neuron has access to information accumulated from sequential information derived from neurons in front of and behind it; these have been shown to be computationally efficient and highly accurate (Graves & Schmidhuber, 2005). Neural networks have shown high accuracy even when sample sizes are small (Saez, Baldominos, & Isasi, 2016). Beyond supervised techniques, neural networks are also used for dimensionality reduction purposes, offering a nonparametric alternative to principal components analysis (Hinton & Salakhutdinov, 2006).

3.4.1.5 Regularized linear regression methods

Regularized linear regression methods are approaches that regularize the coefficients when modeling a continuous outcome. They are especially useful when the predictor set is quite wide and when collinearity is a concern. In essence, these methods (which include ridge regression, least absolute shrinkage and selection operator [LASSO] regression, and elastic net regression; Tibshirani, 1996) shrink the magnitude of coefficients such that only the most influential coefficients remain robust (in this way, regularized linear methods also operate as feature selectors). In terms of the LASSO, which is the method used for RQ1, tuning parameter λ penalizes the regression coefficients, with greater values of λ shrinking coefficients to zero. A cross validation approach can determine the optimal value of λ, which will select the most important variables from the lexical and syntactic speech features as predictors of students' reading comprehension score.

3.4.2 Feature importance

Once the best model was chosen, the importance of each feature was ascertained using the ELI5 Python package (Korobov & Lopuhin, 2019). The method employed was permutation importance, where during each iteration a single variable is permutated, or randomly scrambled. The mean absolute error is calculated, which can be interpreted as the error in the model when

that variable is not available. This process is repeated for each variable in the model, for a specific number of iterations, which in this study was 10. The reduction in mean MAE (and standard deviations) is reported, with larger values indicating a greater loss in predictive power when that variable is missing.

ELI5 can gauge feature importance either by (1) using the entire dataset to inspect the features that are most important in building an existing estimator (less generalizable), or (2) through a cross-validation method where the feature's importance is measured by how well it predicts an unseen test set (more generalizable). Due to the small sample sizes, I used approach (1). It is important to reiterate that while I tested models with an aim for generalizability, that is, through cross-validation, the importance metrics for individual features in our final models are not necessarily generalizable outside these specific models, due to the use of approach (1).

3.5 Research question 2: Interactions between NLP features and demographics in predicting reading comprehension

In RQ2, I explored any interactions the top five features in each of the four top models (oral/regular, oral/modified, text/regular, and text/modified) from RQ1 might have with the two self-reported demographic variables (years in Canada and total language proficiency) in predicting reading comprehension. I used the R package 'ranger' (Wright, Wager, & Probst, 2019) to examine pairwise interactions among these variables. Specifically, I used a version of 'ranger' (1.6.33) reported by Wright, Ziegler, and König (2016) where the authors examined pairwise gene-gene interactions. Variables are permutated in specified pairs (i.e., one predictor variable and one demographic variable), and the output indicates the increase in MAE for the permutation of each pair. Although significance levels are not available for the pairwise permutations, results can be compared to the permutation of each variable individually. Any pair that has a greater MSE value than the permutation of each individual variable in the pair can be considered a candidate for a significant interaction effect. For such candidates, I ran a multiple regression (R Core Stats Package, R Core Team, 2019) to identify and plot where interactions may exist.

3.6 Research question 3: Unsupervised ML to determine latent patterns in NLP and their relationship with reading comprehension

In RQ3 I used an unsupervised ML method, factor analysis, to determine the factor structure underlying the four NLP datasets (oral/regular, oral/modified, text/regular, and text/modified). I then regressed the reading comprehension score on the resulting factors. The aims of this analysis are to understand the pattern of factor loadings of NLP-derived variables, and to determine if any of these factors are predictive of reading comprehension. These results are then compared with the supervised models for commonalities and differences.

As described above, supervised ML configures parameters that model the outcome variable as a function of the independent variables in a training set, with the goal of minimizing error in the outcome variable. Then, this algorithm is evaluated by how accurately its estimated parameters, when applied to the independent variables in the novel test data, predict the outcome of the test data. In unsupervised ML models, however, outcome variables are not considered. Rather, unsupervised algorithms look for patterns that explain variance in data matrix *X*. Unsupervised models can be evaluated in terms of how closely data points cohere within data clusters and how distinctly the clusters separate from one another (Deborah, Baskaran, & Kannan, 2010). Generally, the goal of unsupervised ML is to decompose a high-dimensional matrix of predictor variables into fewer dimensions, while not omitting important information. Unlike feature selection, which was the aim of the recursive feature elimination in some of the supervised approaches described for RQ1, unsupervised models aim for feature reduction, also known as feature extraction or dimensionality reduction. The fundamental idea is that predictor variable matrix *X* can be decomposed into a certain number of factors or components that retain the important variance of the variables in matrix *X*. Clustering methods such as *k*-means clustering, principal component analysis, and factor analysis are typical approaches to unsupervised ML.

This study uses factor analysis to identify continuous latent factors in the independent data matrix, that is, the NLP-derived lexical and syntactic features. The analysis was performed using Scikit-learn's FactorAnalysis (Pedregosa et al., 2011), which assumes that a set of latent factors generate the manifest, measurable set of NLP-derived features. Factors are linear combinations of variables that have high correlations (either positive or negative) amongst themselves and low correlations with other factors. The algorithm identifies the latent factors that maximize the

shared, or common, variance among features. Variance not shared by features loading onto a factor is either variance specific to that feature, or it is variance due to measurement or human error. While principal component analysis is a typical unsupervised approach in ML, one of its assumptions is that the error variance across all features is equivalent. This assumption cannot hold in the current study, so instead factor analysis is used, because it can model both specific and error variance. Factor analysis in Scikit-learn performs an iterative maximum likelihood estimation of the factor loading matrix. It differs from typical exploratory factor analysis in that it utilizes expectation maximization, or EM (Pedregosa et al., 2011; Poznyak, 2018). While maximum likelihood (ML) seeks to estimate the parameters that maximize the likelihood of producing the given data, factor analysis seeks to unearth latent, or hidden, factors underlying the given data, as well as the parameters. Since there is no initial guidance for the algorithm to identify the parameters, EM makes an initial guess at the distribution of the latent factors based on the observed data (expectation step), which are then evaluated and recalculated to maximize the parameter values (maximization step). Analogous to gradient descent, mentioned above, this process repeats until convergence occurs, i.e., a unique solution (global optimum) of the maximum likelihood function is reached.

In turn, the factors identified in the unsupervised model can be used in supervised ML models, or, as in this study, as variables in traditional inferential statistics. Specifically, I entered the lexical and syntactic factors for each of the four datasets (oral/regular, oral/modified, text/regular, and text/modified) into multiple regression equations with the reading comprehension score as the outcome variable. This process was necessary because the high dimensionality of X (wherein the number of variables is greater than the number of observations) would have precluded model identification in a traditional multiple regression model.

I then compared the results from this research question to the results of RQ1. These two analyses differ in that RQ1 models reading comprehension using the NLP-derived lexical and syntactic dataset as a "bag of features" – meaning that the individual features' contributions to predicting the reading comprehension outcome are additively modeled (i.e., removing one feature from the model does not necessarily change the structure of the model – it just somewhat reduces accuracy, as demonstrated by the permutation process for calculating feature importance in RQ1). The relationships *among* the lexical and syntactic features themselves is not of particular interest in RQ1; the focus instead is on how each predicts reading comprehension. In RQ3,

however, the latent structure of the lexical and syntactic features is of prime interest: which lexical and syntactic features load onto a common factor, if any? Do these commonalities differ across the four datasets (oral/regular, oral/modified, text/regular, and text/modified)? In other words, if the first modelling step ignores the reading comprehension outcome, and allows the lexical and syntactic factors to emerge in an unsupervised manner from the oral-elicited and text-elicited datasets, do the resulting factors have a predictive relationship with reading comprehension (when modeled using multiple regression)? Is a similar amount of variance in reading comprehension explained by the latent factors that underlie the lexical and syntactic feature sets as when the "bag of features" (RQ1) supervised method is used?

4 Findings

4.1 RQ1: Supervised model-building and feature importance

The first research question asks: how much variance in reading comprehension can be modelled as a function of individual productive lexical and syntactic features extracted through NLP? How does the variance explained compare to that explained by published models that use traditional lexis and syntax measures? What are the top lexical and syntactic feature predictors for each of the four datasets (oral/text elicitation *by* more/less skilled readers)? How do the top predictors differ across the four models?

Pairwise correlations were found between each of the NLP-derived linguistic features and reading comprehension score, for each of the four datasets: oral- and text-elicited, regular and modified versions of the reading comprehension assessment. Due to length, the correlation table (Table 34) can be found in Appendix A.

4.1.1 Descriptive correlations between individual lexical and syntactic features and reading comprehension outcomes

There are several syntactic and lexical features that show consistent positive or negative correlations across the four datasets. The descriptive correlations (Appendix A) show consistent positive trends between grammatical features (i.e., constituents or parts of speech) and reading comprehension for adjectives, phrases with prepositions, subordinating conjunctions, and wh-noun phrases. That preposition use correlates with reading comprehension is expected given the documented contribution of prepositions to writing quality (Crossley et al., 2016). The positive correlation between reading comprehension and features relating complex grammar (subordinating conjunctions) was also expected. Negative correlational trends across all four datasets included noun phrases that consist of noun phrase, coordinating conjunction, and noun phrase (which may be expected due to the lack of cognitive demand that coordinating conjunctions such as “and” require), and verb phrases consisting of a non-3rd person singular present verb and noun phrase.

In terms of grammatical complexity, positive correlational trends consistent across all four datasets include the average length of verb phrases, mean length of sentences, number of verb phrases per T-unit, number of prepositional phrases (over the length of each sample), and the

average height of each parsed sentence tree in the sample (and sum of tree parse depths for all words in each sentence). These grammatical complexity features align with the existing knowledge base around use of complex grammar and literacy attainment. Negative trends exist across the four datasets for coordinate phrases (e.g., 'and'), which aligns with the grammatical constituent findings above, in that phrases with coordinating conjunctions were negative predictors in that category. A T-unit is the minimally terminable unit of grammar (for example, a clause with many subordinate clauses can be a T-unit because the subordinate clauses cannot stand on their own, but not two clauses connected with a coordinate conjunction is not a T-unit because the clauses are not subordinated and they can stand on their own). The number of T-units (normalized by sample length) was also consistently negative, suggesting that a greater number of T-units is a trade-off for grammatical complexity. This inverse relationship is demonstrated in Figure 9.

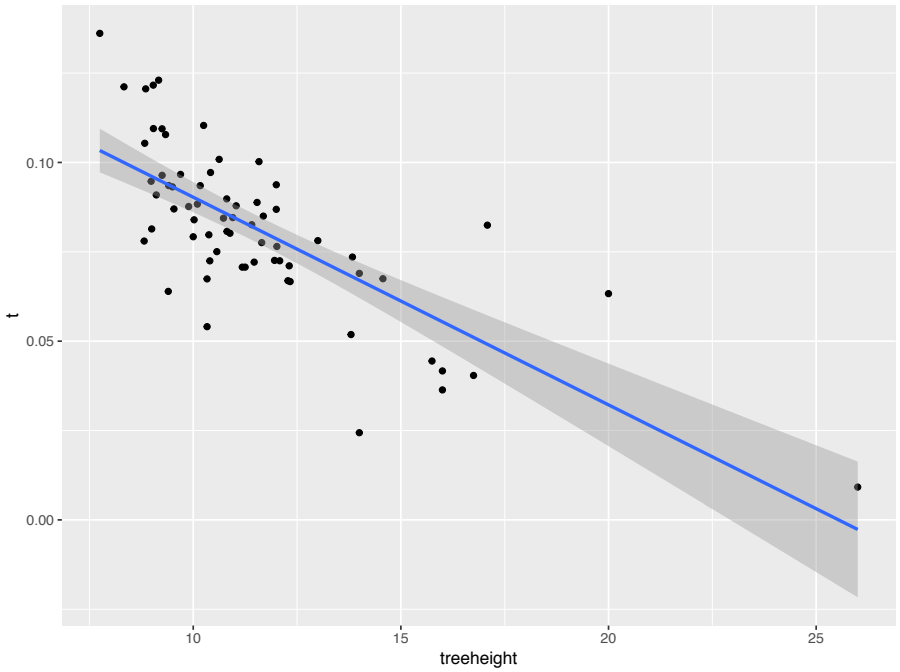

Figure 9. Scatterplot of average parsed sentence tree height (x-axis) by number of T-units normalized by sample length (y-axis)

As for vocabulary range, consistently positive correlations with reading comprehension across the four datasets include average length of words and age of acquisition of verbs, which are expected because more skilled readers are likely to use more sophisticated vocabulary. Negative trends existed for imageability (of words, nouns, and verbs), not-in-dictionary words, and verb

frequency. These correlations are expected, as higher scores in imageability and frequency are associated with less uncommon and sophisticated vocabulary with higher scores. No vocabulary richness feature (i.e., type-token ratio, or moving average type-token ratio) was consistently positive or negative in all models.

In word specificity, similarity, and ambiguity, consistent positive correlations with reading comprehension were found with the standard deviations in the longest (and shortest) path from a given word to its hypernym root; this is a measure of variation in vocabulary specificity. On the other hand, the average of such paths, which would indicate average specificity of vocabulary, were consistently negatively correlated with reading comprehension, which might be considered an unexpected finding. However, as discussed earlier, the word specificity as operationalized in WordNet is based on semantic specificity, rather than lexical sophistication. Figures 10 and 11 demonstrate the lack of strong relationship (or perhaps slightly negative) between the average maximum path from a given word to its hypernym (specificity) and word length, as well as by age of acquisition. These figures indicate that it cannot be assumed that there is a positive association between the WordNet specificity metric and the subjective or objective vocabulary metrics. For contrast, Figure 12 shows the stronger positive relationship between age of acquisition and word length. Thus, results that include WordNet specificity should be interpreted with caution.

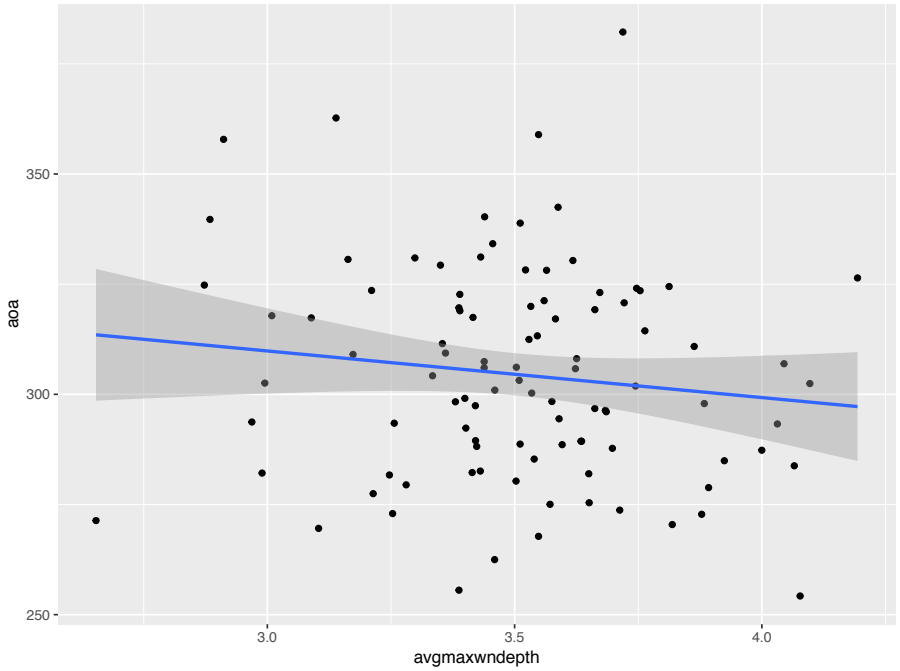

Figure 10. Average maximum depth from a given word to its root hypernym (x-axis) by age of acquisition (y-axis) in the text/regular dataset

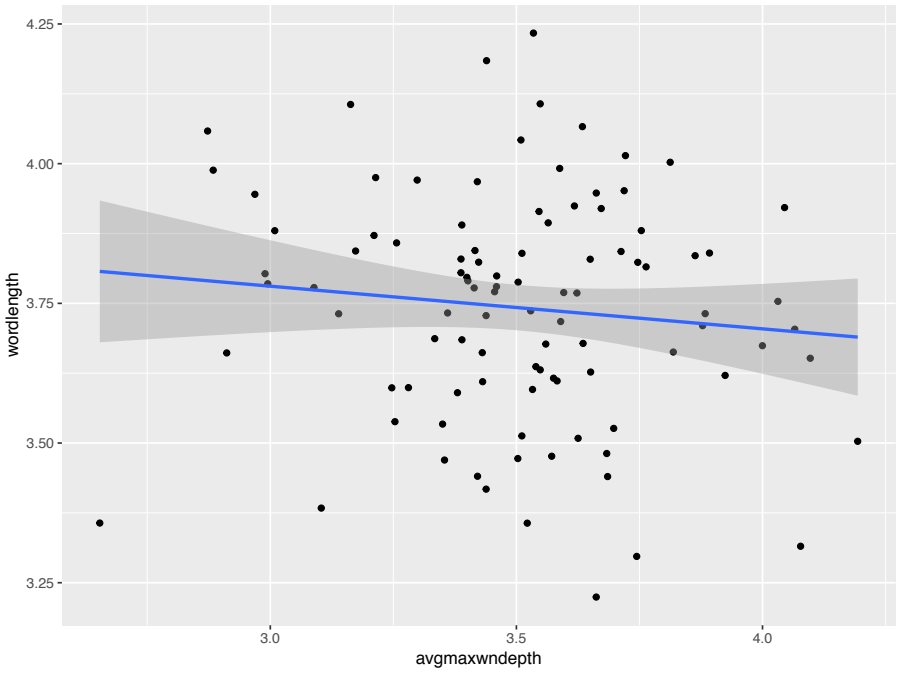

Figure 11. Average maximum depth from a given word to its root hypernym (x-axis) by word length (y-axis) in the text/regular dataset

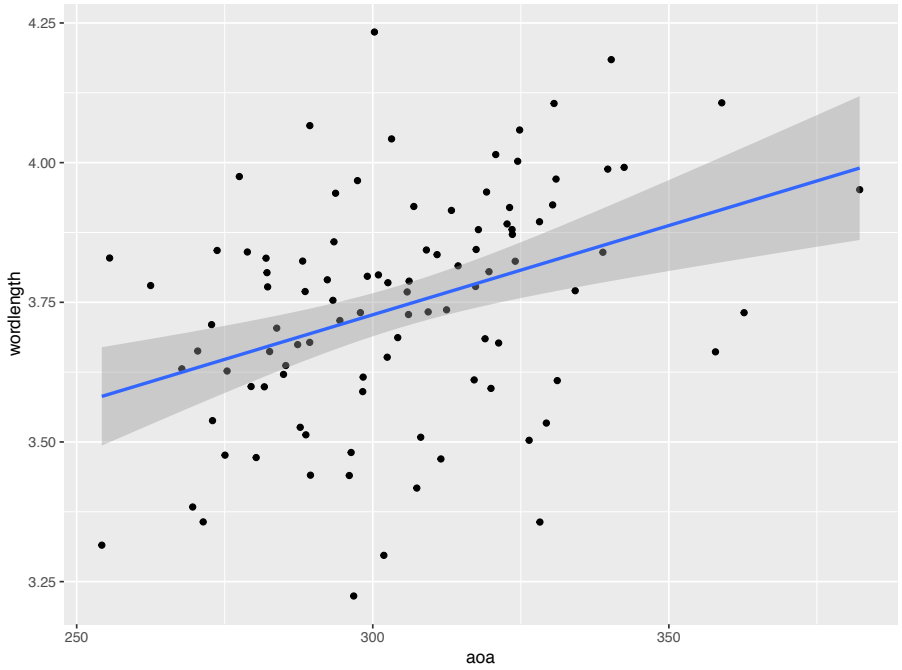

Figure 12. Age of acquisition (x-axis) by word length (y-axis) in the text/regular dataset

The average word meaning similarity for three of six WordNet methods showed consistently negative correlations with reading comprehension (but the other three methods had quite divergent results). This could relate to repetitiveness, which may be associated with less deep cognitive processing.

Two sentiment and affect features had consistent trends: the MPQA strong negative was a consistently positive predictor of reading comprehension, while the mean Stanford very positive sentiment was a consistently negatively correlated with reading comprehension.

4.1.2 Machine learning model results

Results from the ML model building and test are presented in Table 6. Recursive feature elimination with cross-validation (RFECV) was used for preprocessing in all but the support vector model (for which it is not available).

Table 6

Comparison of mean absolute error (MAE, SD in parentheses) for four regression-based ML models predicting reading comprehension from lexical and syntactic linguistic features

	Mean Baseline	Gradient Boosting	K-Nearest Neighbors	Decision Tree	Support Vector	Random Forest	LSV	MLP	Lasso

SPEECH (ORAL) ELICITATION									
Regular version of reading comprehension measure									
Default	10.21	10.39 (1.42)	10.64 (1.47)	13.97 (1.92)	10.05 (1.47)	9.90 (1.61)	82.44 (9.60)	33.41 (4.65)	10.31 (1.32)
Tuned		9.96 (1.46)	10.49 (1.48)	9.77 (1.58)	**9.59 (1.31)**	9.70 (1.32)	69.17 (10.33)	17.45 (5.43)	10.21 (1.33)
Modified version of reading comprehension measure									
Default	10.30	8.18 (1.89)	10.20 (1.88)	13.26 (2.86)	10.29 (1.83)	9.29 (1.78)	135.53 (19.45)	46.71 (5.19)	10.54 (1.90)
Tuned		**8.12 (1.88)**	9.72 (1.71)	12.65 (2.52)	9.94 (1.83)	8.76 (1.80)	85.20 (3.33)	17.27 (2.11)	10.30 (1.87)
TEXT (WRITING) ELICITATION									
Regular version of reading comprehension measure									
Default	9.96	8.83 (1.33)	9.93 (1.75)	12.22 (1.98)	9.67 (1.52)	9.29 (1.37)	82.15 (10.40)	41.70 (4.32)	9.56 (1.54)
Tuned		8.71 (1.41)	9.10 (1.59)	11.81 (1.96)	8.98 (1.38)	**8.69 (1.41)**	71.36 (11.59)	37.42 (6.84)	9.34 (1.50)
Modified version of reading comprehension measure									
Default	10.41	10.82 (1.89)	11.23 (1.95)	14.41 (2.69)	10.51 (1.99)	10.76 (1.75)	139.41 (15.25)	45.04 (6.39)	10.55 (1.98)
Tuned		10.72 (1.91)	10.46 (1.87)	13.80 (2.42)	10.39 (1.90)	**10.16 (1.96)**	85.55 (3.86)	17.30 (2.87)	10.41 (1.86)

Note: Selected models for each prediction problem are bolded. LSV: Linear support vector; MLP: Multi-layer perceptron.

Tuned models outperformed default parameters in all cases. Random forest regression was the best model for both text-elicited datasets. The best model for the regular/oral elicitation was support vector regression, while gradient boosting regression best predicted the oral/modified elicitation. Table 7 summarizes the functioning of the best four models.

Relative error reduction is calculated by finding the difference between the mean baseline MAE and the best tuned model MAE and dividing this by the mean baseline MAE. The model that had the best relative improvement from the mean baseline was the oral-elicited modified dataset with a 21.17% improvement, with the next-best the text-elicited regular (12.75% improvement),

followed by oral-elicited regular (6.07% improvement) and text-elicited modified (2.40% improvement). Differences in relative error reduction across the four models can be attributed to the varying strength in the predictive relationship between the reading comprehension and the set of NLP-derived lexical and syntactic features. In addition, outliers can influence the baseline model (which uses the mean of the outcome variable to train the model, rather than the actual scores), and if those outliers do not have a strong relationship with the dependent variables, then they will continue to exert an effect even with the best model.

The R^2 values in the final column in Table 7 are not simply the total reading comprehension variance explained by the set of syntactic and lexical features in the entire dataset, as R^2 usually refers to in traditional inferential statistics. Rather, R^2 here refers to the variance explained only in the testing data. This is calculated just as typical R^2, by squaring the sum of the difference between each predicted outcome ($\hat{y}$) and actual outcome (y) for each point in the test data (i.e., residuals), and dividing this by the total variance in outcome value; then this value is subtracted from 1; see equation below from the Scikit-learn documentation.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

The difference between R^2 in traditional regression models and R^2 as a ML metric is that the former represents the total variance in the outcome variable explained by the set of independent variables; is a metric that describes how well a function of input variables can explain an outcome variable. The R^2 metric in ML, on the other hand, evaluates how well the function can be applied to unseen data. It describes how well the algorithm has learned the relationship between the independent variables (linguistic features) and dependent variable (reading comprehension scores) in the training dataset and how well it can apply those learnings to a novel test set.

Table 7

Summary of best models for oral- and text-elicited, regular and modified versions

Dataset	*n* size (train size, test size for CV)	RFE with CV result	ML model	Baseline CV MAE (SD)	Tuned CV MAE (SD)	Relative error reduction (CV)	R^2 on individual train/test split
Oral-elicited, regular	95 (76, 19)	n/a	SVR	10.21 (1.33)	9.59 (1.31)	6.07%	20.4 (Figure 13)
Oral-elicited, modified	70 (56, 14)	43 of 260 variables	GBR	10.30 (1.87)	8.12 (1.88)	21.17%	22.4 (Figure 14)
Text-elicited, regular	99 (79, 21)	211 of 260 variables	RF	9.96 (1.47)	8.69 (1.41)	12.75%	36.6 (Figure 15)
Text-elicited, modified	67 (54, 13)	87 of 260 variables	RF	10.41 (1.86)	10.16 (1.66)	2.40%	18.5 (Figure 16)

Note: CV: cross-validated; RFE: recursive feature elimination; MAE: mean absolute error; SD: standard deviation; R2: percent of variance explained; SVR: support vector regression; GCR: gradient boosting regression; RF: random forest

Cross validation was used to find the best model for each of the four datasets. For visualization purposes and to compare to existing literature, the R^2 values in Table 7 were developed by creating a single 80/20 train/test data split, running the tuned model on the training data, and then testing the model's predictions on the test data. The R^2 values reported in the last column of Table 7 are not as reliable as the cross-validated model information provided in the form of MAE in columns 5 and 6, since the cross-validation process averages predictive power over 200 different train-test splits. The difference is that the R^2 value represents a single train/test split while the relative error reduction is ascertained through cross-validation, that is, the average reduction in error over 200 different train/test splits. For example, the best R^2 value is for the text/regular dataset, while this dataset did not exhibit the best cross-validated relative error reduction.

When creating a training/test partition, the choice of random seed influences which datapoints will be in the training set, and which will be in the test set. As mentioned above, if high-leverage observations are located in the training set, this can impact how well the model can predict the unseen test set, and vice versa if they were located in the test set. An outlier is likely to exert more influence when in a 20% test set than when part of an 80% training set. In other words, the

R^2 values for a single train-test split can be unstable. This explanation addresses the discrepancies in Table 7 between the best functioning model in terms of relative error reduction modeled through cross-validation and R^2 on a single train/test split. With larger datasets, there would be smaller discrepancies between the two metrics. These data were analyzed for high-leverage observations and only those values below .50 correct on BALA were removed. Nonetheless, the linguistic features may have some high-leverage data points; in addition, many of them are counts normalized by length rather than normal distributions, which could introduce instability.

Figures 16 to 19 are visual representations of the R^2 just described, found in the last column of Table 7. As these plots indicate, the algorithm did not do well predicting the entire range of the actual outcome variable, as the distribution along the y-axis ($\hat{y}$) is truncated. The results that explain relatively more variance (text-elicited/ regular, and oral-elicited/modified) have the least truncated y-axes, the R^2 values are higher, and the slope of the fit line is steeper.

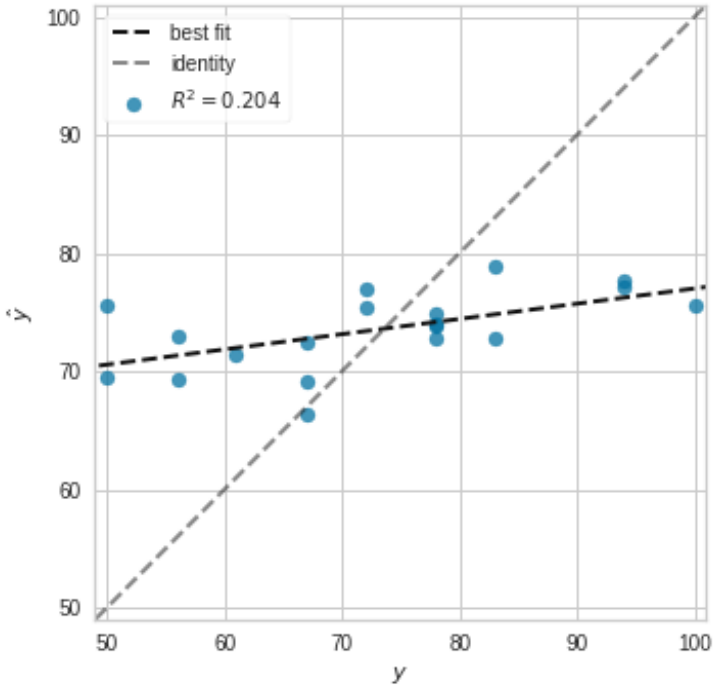

Figure 13. Predicted ($\hat{y}$) by actual (y) outcomes for a single train/test split, oral-elicited, regular version of reading comprehension assessment (support vector regression model, n=95)

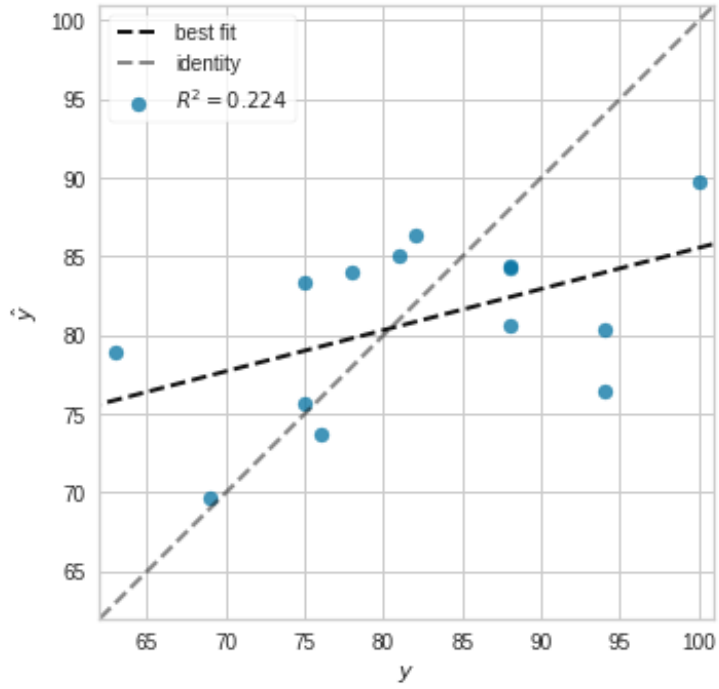

Figure 14. Predicted ($\hat{y}$) by actual (y) outcomes for a single train/test split, oral-elicited, modified version of reading comprehension assessment (gradient boosting regression model, n=70)

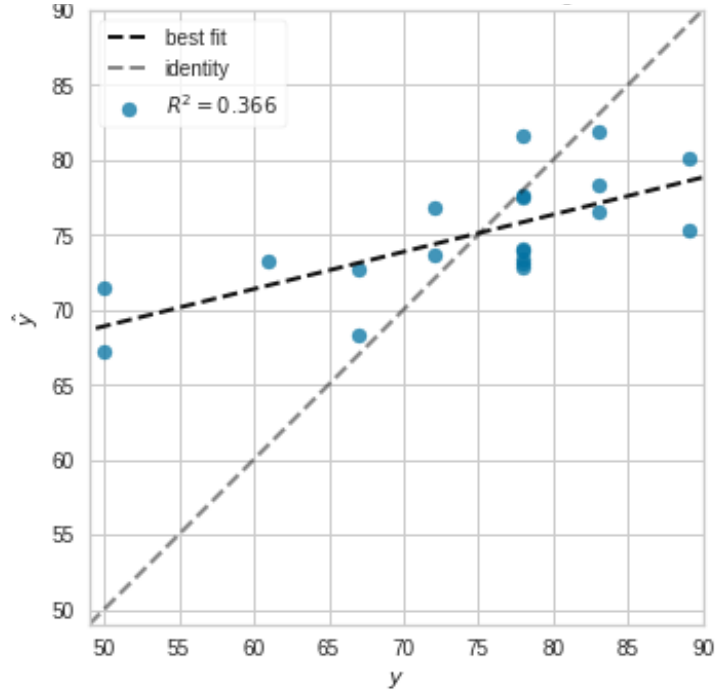

Figure 15. Predicted ($\hat{y}$) by actual (y) outcomes for a single train/test split, text-elicited, regular version of reading comprehension assessment (random forest regression model, n=99)

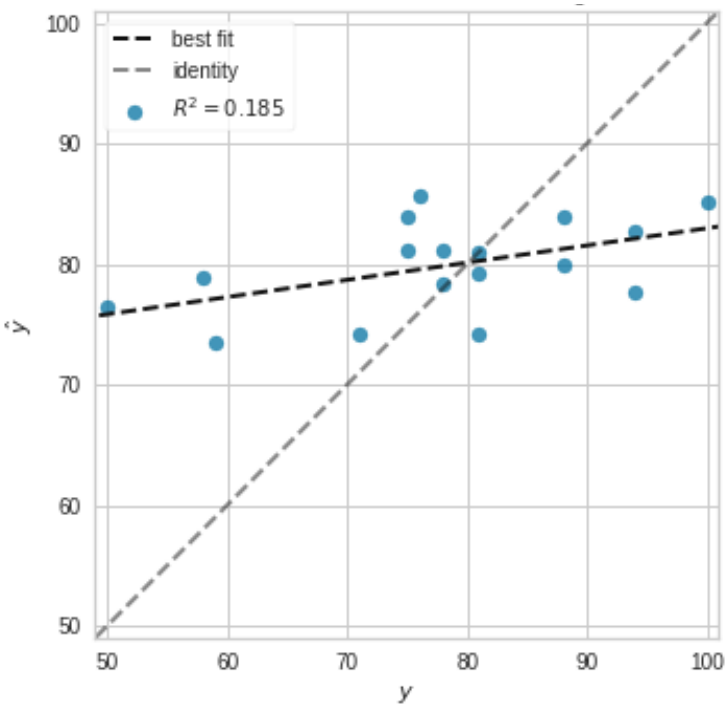

Figure 16. Predicted ($\hat{y}$) by actual (y) outcomes for a single train/test split, text-elicited, modified version of reading comprehension assessment (random forest regression model, n=67)

4.1.3 Feature importance results

The top features of the best models of each of the four datasets were ascertained through random permutation. After the best model was selected for each of the four datasets, each variable was randomly permutated individually while the model was run, to determine the decrease in mean absolute accuracy (MAE) with that variable randomized.

Tables 8-11 present the top twenty features for each model. The MAE column represents the increase in mean absolute error when that variable is permutated. Direction refers to whether the linguistic feature is positively or negatively predicting reading comprehension. Positive predictions correlate positively (higher values are associated with higher reading comprehension scores); while negative predictions operate in reverse. The examples for each predictive feature (last column) were selected from among the three oral-elicited (picture description or story retell) or text-elicited (writing or open reading response) samples that had the highest scores for that feature. Due to model specifications and perhaps a limitation of the ELI5 package, the increases in MAE for permuted features in the first model (the oral-elicited, regular dataset using support vector regression) are substantially smaller, relatively speaking, than those in the random forest

and gradient boost regression models. Nonetheless, the features represent the top 20 predictors in that model.

As an overview, Figure 17 indicates the distribution of the five different linguistic feature types (grammatical constituents, grammatical complexity, vocabulary richness and range, vocabulary specificity, similarity, and ambiguity, and vocabulary affect and sentiment) across the 20 top predictors for each of the four datasets (oral/regular, oral/modified, text/regular, and text/modified). For all four datasets, the grammatical constituents were the most common top predictors, and they were more likely to be in the top 20 predictors for the regular datasets than the modified datasets. Grammatical complexity features were the least common type to be found in the top 20 predictors. Vocabulary richness and range were only slightly more likely than grammatical complexity to be in the top 20 predictors, and they were more likely to be found in the top 20 features of the regular models than modified models. For the most part, vocabulary specificity, similarity, and ambiguity were also not common in the top 20, except for the text-modified dataset, where 7 features of this type were identified as top predictors. Affect and sentiment predicted reading comprehension more strongly in the modified datasets than the regular datasets.

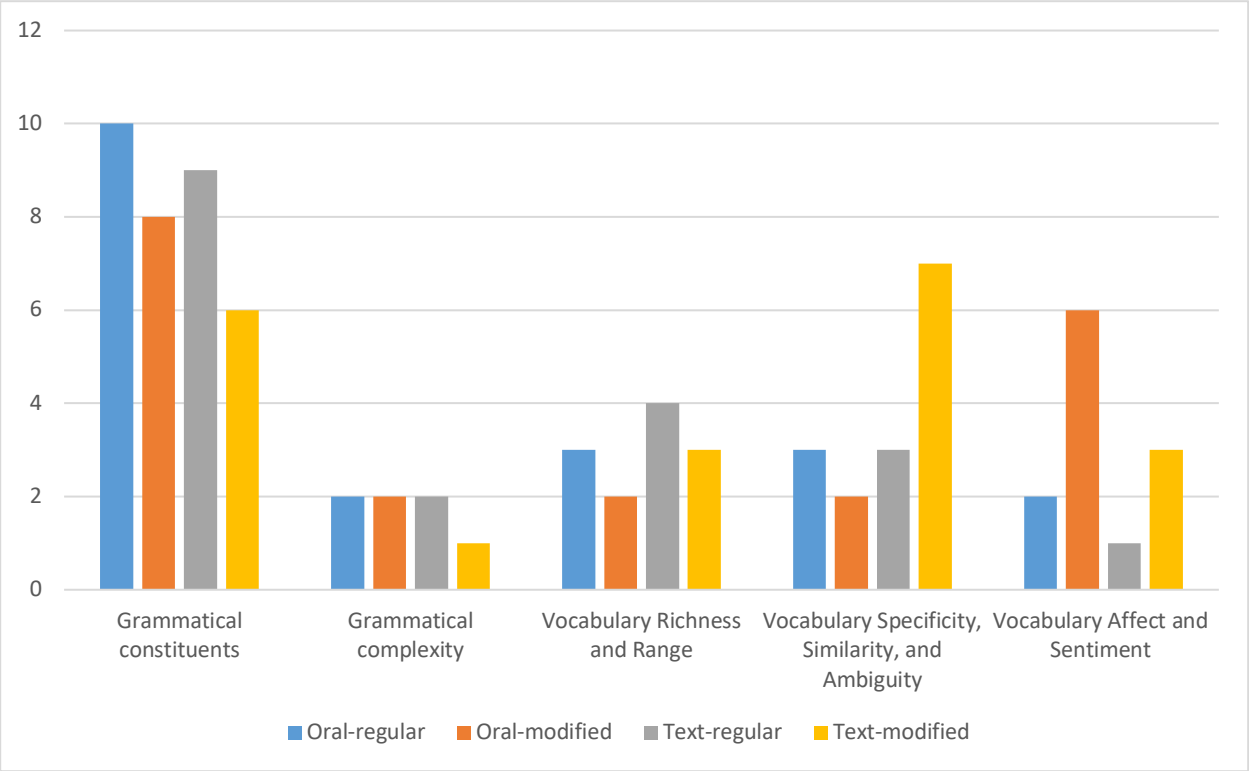

Figure 17. Distribution of top 20 lexical and syntactic feature predictors for the four datasets by type of feature

Table 8

Twenty most important NLP-extracted grammar and vocabulary features in the oral-elicited/regular dataset, predicted through support vector regression permutation (n=95)

ID	Feature (source)	Feature type	MAE (Dir.)	SD	Example student response with high indicator for the feature
OR1	Content density: proportion of nouns, verbs, adjectives, and adverbs (Roark)	Gram-const	.08 (+)	.07	"…Mac mumbled I've always wanted to be on the school track and field team but I was too nervous to ask so then they all race against each other and they all made it just as [um] the recess bell rang a week later they all joined the team" (P847, SR2)
OR2	Prepositional phrase consisting of a single prepositional phrase or subordinating conjunction (Stanford POS tagger)*	Gram-const	.07 (−)	.06	"*on* her first day *of* school she woke up *at* six thirty am and ate breakfast right *after* that she missed the bus so she had to ask her dad to drive her *to* school but the car needed gas so they had to go *to* the gas station to fill the car up *with* gas …" (P775, SR1)

OR3	Noun phrase: noun phrase → coordinating conjunction → noun phrase (Stanford POS tagger)*	Gram-const	.05 (−)	.05	"I see a police officer falling off a bike *and* I see a pink car a dog *and* a cat *and* a person in the back running" (P891, PD2
OR4	Noun phrase: determiner → plural noun (Stanford POS tagger)*	Gram-const	.05 (+)	.05	"the mom is making food for *the kids* and *the kids* are helping out with the cooking…the boy is playing with the dog" (P851, PD1)
OR5	Verb phrase: non-3rd person singular present verb (Stanford POS tagger)*	Gram-const	.05 (+)	.06	"Suzanne *wake* up early for first day for her school" (P797, SR1)
OR6	Variation in word specificity: maximum depth from each verb to root (WordNet)	Voc Spec-Sim-Amb	.05 (+)	.06	"One day at recess Pablo and Ariana were *racing* … they were *determined* to make it onto the track team … as Pablo was *arriving* the to the finish line a big tall figure *was* in his path … "Get out of my path," *grumbled* Mac…" (P859, SR2)
OR7	Strong negative sentiment of words (MPQA)	Voc Aff-Sent	.05 (+)	.04	"Pablo started running when the school bully stood in front of him and he was too scared to talk to the bully [um] then Ariana came and she helped Pablo tell off the bully [um] and then Pablo asked if mac the bully wanted to race with them" (P928, SR2)
OR8	Count of verbs (Stanford POS tagger)*	Gram-const	.04 (+)	.05	"It *was* the first day of school she *went* downstairs *made* herself a bowl of cereal and *started eating*…her dad *gave* her a big hug and *said* goodbye she *got* her shoes on *grabbed* her bag and *ran* out the door to *catch* the bus when the bus *starting coming* around the corner Suzanne *realized* she *forgot* her lunch" (P823, SR1)
OR9	Noun phrase: Noun phrase → verb phrase (Stanford POS tagger)*	Gram-const	.04 (−)	.06	"In this picture *I see a girl falling off her bike* and two *animals jumping* out of the way" (P921, PD2)
OR10	Variation in word specificity: minimum depth from each verb to its root (WordNet)	Voc Spec-Sim-Amb	.04 (+)	.05	"they *wanted* to be in the track and field team so they *practiced* really hard every day and there was a student *called* mac who always *bullied* Pablo" (P825, SR2)
OR11	Complex nominal per clause (Lu)	Gram-com	.04 (−)	.05	"There is a *coffee store* in the back with a pink car with some jogging two kids walking *police woman* falling off her bike and a dog and a cat jumping across the *bike lane*" (P946, PD2)

OR12	Verb phrase: gerund/present participle verb → noun phrase → prepositional phrase	Gram-const	.04 (+)	.05	"Three people are in the kitchen [uh] they are a family and so the mom is *making juice or smoothies with the blender*" (P803, PD1)
OR13	Coordinate phrases per clause (Lu)	Gram-com	.04 (−)	.05	"…He saw a bully who was trying to pick on him... Pablo's friend came *and* stood up to him *and* asked if he wanted to practice with them…they became friends *and* made the track team (P945, SR2)
OR14	Honoré's statistic	Voc Rich-Range	.04 (+)	.03	"Suzanne woke up on the first day of school on September eight she had cereal for breakfast and she put on her shoes and went out the door…" (P746, SR1)
OR15	Word imageability (Bristol and Gilhooly-Logie)	Voc Rich-Range	.04 (−)	.04	"A person is riding and two woman are like walking a boy is going to fall down off the bike a dog is chasing after a cat and a car a pink car…" (P913, PD2)
OR16	The mean frequency of each verb's occurrence in a lexical corpus (Brysbaert & New)	Voc Rich-Range	.03 (−)	.03	"…the daughter *wants* a cupcake and the cupcake might *fall* over and the toaster [um] *has* bread … the son *has* a dog in the kitchen *playing* with a ball and [uh] the mother is *putting* apples in … the blender (P772, PD1)
OR17	Simple declarative clause: verb phrase (Stanford POS tagger)*	Gram-const	.03 (+)	.04	"One kid's *playing* with his dog and another *is trying to get* [um] cupcakes on top of a fridge *using* a stool that *is* about *to fall* and the mom *looks* like she's *blending* something and *trying* to *grab* an apple" (P872, PD1)
OR18	Count of adverbs (Stanford POS tagger)*	Gram-const	.03 (+)	.05	[Mac said] "I'm *too* nervous… they *eventually* realized that Mac was *also* scared like them …they *eventually* tried out for the track and field team and *later* days *later* they made the team (P888, SR2)
OR19	Negative word sentiment (Stanford sentiment analysis)	Voc Aff-Sent	.03 (+)	.03	"[Suzanne] felt embarrassed she had to ask her dad to drive her to school…her dad said the gas was low so they had to stop and get some gas Suzanne felt embarrassed on her first day of school she was seven minutes late (P857, SR1)
OR20	Variation in word similarity (WordNet Resnick SemCor method)	Voc Spec-Sim-Amb	.03 (+)	.05	"…the girl is standing on a stool trying to grab some cupcakes and the stool is tipping there's a mixer with some

	carrots and a banana peel on top there are two clocks hanging from the shelf the radio is playing music" (P727, PD2)

Note: MAE = mean absolute error; SD = standard deviation; Dir.=direction of prediction (positive or negative predictor of reading comprehension outcome). * normalized by total number of words in each response.

The top five features in the oral/regular model were grammatical constituents. The top predictor, content density, is the ratio of "open class" parts of speech (those containing a virtually unlimited number of words, including nouns, verbs, adjectives, and adverbs) to "closed class" words which have a finite number of words, such as prepositions, determiners, and conjunctions. High content density potentially indicates that the language produced contains more meaningful information than low content density. Count of certain noun phrases (OR4, OR9), verb phrases (OR5, OR8, OR12, OR17), and adverbs (OR18), accordingly, were positive predictors. Because the oral tasks (picture description and story retelling) require the participant to narrate actions, the use of action verbs is essential. Use of verbs also would impede the tendency to list objects or entities when narrating. That simple prepositional phrases (OR2) are negative predictors aligns well with content density (OR1), because prepositions are in the denominator of content density. Similarly, the negative predictive power of coordinating conjunctions (OR3, OR13) suggests that extensive use of "and" may result in longer, but not more complex, sentences.

The negative prediction of complex nominals (OR11) is somewhat less expected but may relate again to the use of listing, or perhaps of formulaic language. An example of an oral/regular narrative with a high score for complex nominals: "there is a *dog with their red ball* … one mother … blending maybe a *smoothie of some sort* … there is a *toaster with toast ready* …there is a grey stove and a *white counter top with purple and light blue pulls* …" (P915, PD1). While not easily interpretable, the use of complex nominals could relate to the "listing" type of language that does not appear to actively connect the propositions in the heard/read/seen prompt, but instead just recites a list.

Vocabulary richness and range do not emerge as important predictors until OR14-16, where they behave as expected: Honoré's statistic (OR14), which values words used only one time in the entire narrative, is positively associated with the reading outcome, while word imageability

(OR15) and verb frequency (OR16), increases in which represent more concrete and common words, are negatively associated with it.

Three vocabulary specificity, similarity, and ambiguity features were top predictors positively associated with the reading outcome: variation in the maximum (OR6) and minimum (OR10) depths from a word to its hypernym (this is a variation in word specificity), and variation in word similarity (OR20). These suggest that using a mix of less and more specific words in the oral narratives was more predictive of reading success than simply using highly specific words. As for the variation in word similarity, this is positively predicting reading possibly because a mix of some similar words and some dissimilar words may allow for more cohesive narrative. Finally, two similar affect and sentiment variables emerged as top predictors: the MPQA strong negative sentiment (OR7), and the Stanford sentiment analysis negative word sentiment (OR19), which refers to the proportion of words in the participants' narratives that are tagged as "strong negative" or "negative" in these corpora. The four oral elicitation tasks did have dramatic elements (a child missing the bus and arriving late for school, a playground bully, a child about to fall from a stool, and a police officer falling from her bike), therefore, an accurate retelling or description would involve using some words with negative sentiment.

Absent from the top predictive features in this model were features relating to type-token ratio, the arousal, dominance, and valence of vocabulary, word ambiguity, and any positive predictors for grammatical complexity.

Table 9

Twenty most important NLP-extracted grammar and vocabulary features in the oral-elicited/modified dataset, predicted through gradient boosting regression permutation (n=70)

ID	Feature (source)	Feature type	MAE (Dir.)	SD	Example responses from among top three language samples for the feature
OM1	Mean arousal of nouns, from calm to excited (Warriner)	Voc Aff-Sent	2.78 (−)	.62	"This guy Mac he's like the school *bully*…and Pablo was scared of him…when Mac stopped him [he said] 'hey *loser* get out of my face' and then Ariana was like…'he's no *loser'*…" (P737, SR2)
OM2	Verb phrase: "to" → verb phrase (Stanford POS tagger)*	Gram-const	1.93 (+)	.49	"Mac said he wanted *to be* on the team but he was too nervous *to try*… Ariana said 'you can join us *to make* it on the track and field team' … and every recess they

					were running *to make* it on the team…"(P944, SR2)
OM3	Verb phrase: 3rd person singular present verb → verb phrase (Stanford POS tagger)*	Gram-const	1.48 (+)	.41	"… a cat *is running* away from the dog because the dog *is chasing* the cat (P743, PD2)
OM4	Mean arousal of verbs, from calm to excited (Warriner)	Voc Aff-Sent	1.45 (−)	.29	"…she *forgot* her lunch and *ran* back inside the kitchen [and] *grabbed* her lunch from the counter but when she *ran* outside … the bus already *drove* away" (P933, SR1)
OM5	Verb phrase: non-3rd person singular present verb → subordinating conjunction (Stanford POS tagger)*	Gram-const	1.28 (−)	.27	"the little brother is playing with the dog *fetch while* the mom and the sister is getting ready for the picnic" (P755, PD1)
OM6	Noun phrase: singular or mass noun → singular or mass noun*	Gram-const	1.24 (+)	.30	"Suzanne [um] went to the … *school bus* [um] but …she ran back to the *kitchen counter* to get her lunch… She had to stop at the *gas station*" (P949, SR1)
OM7	Mean dominance of nouns from controlled to in control (Warriner)	Voc Aff-Sent	1.18 (−)	.42	"It's a *boy* …and the she *dog* playing *ball* and then the mother make *breakfast* and the listen *music* and the she *sister* put *cake* in *table*" (P918, PD1)
OM8	Noun phrase: determiner → plural noun (Stanford POS tagger)*	Gram-const	1.17 (−)	.36	"Mom is busy by multitasking and working while *the children* are playing with *the dogs* or trying to *some cupcakes*…" (P806, PD1)
OM9	Mean word frequency (Brysbaert & New)	Voc Rich-Range	1.05 (−)	.29	"The police officer fell…and the dog was chasing the cat and in the background there's a man running and two people walking and a car…" (P953, PD2)
OM10	Variation in valence of nouns, from unpleasant to pleasant (Warriner)	Voc Aff-Sent	1.01 (−)	.30	"Pablo and her *friend* was racing on the *track* at *recess* they wanted to get on the *team* but when Pablo was going to reach the finish *line* [um] Mac who was a *bully* [uh] stepped in front of him and blocked him her *friend*" (P788, SR2)
OM11	Dependent clauses per clause (Lu)*	Gram-com	.90 (+)	.21	"She is trying to [um] get the cupcake on top of the fridge except that she is not high enough" (P766, PD1)
OM12	Length of prepositional phrases (Fraser, 2016)*	Gram-com	.62 (+)	.15	"…the music is *on near the audio thingy* the audio music is *on near the apple basket*" (P732, PD1)

OM13	Noun phrase: noun phrase → prepositional phrase (Stanford POS tagger)*	Gram-const	.59 (−)	.20	"Mother is trying to make *juices for the children* while the little girl's trying to take the *cupcake off the fridge*" (P933, PD1)
OM14	Verb familiarity (Bristol and Gilhooly-Logie)	Voc Rich-Range	.53 (−)	.12	"…she *forgot* her lunch [er] she *went* to her house quickly *to get* her lunch but when she *came* the bus *had got* away so she *had to ask* her father *to drive* her to school but the car *was* out of gas so [uh] she [uh] they had to *fill* [uh] it with gas…" (P843, SR1)
OM15	*Wh*-adverb phrase (Stanford POS tagger)*	Gram-const	.48 (−)	.11	"Pablo … was practicing with his friend…*when when* [uh] soon before the recess bell rang *when* he was almost at the finish line then he noticed something [uh] and he stopped *when* he saw his friend was coming" (P761, SR2)
OM16	Strong negative sentiment of words (MPQA)	Voc Aff-Sent	.43 (+)	.09	"There was a girl named Suzanne [um] she was late for school and her dad had to drive her [uh] because she forgot her lunch bag she missed her bus to go to the school and so she was late" (P784, SR1)
OM17	Variation in word specificity: maximum path from each word to its root (WordNet)	Voc Spec-Sim-Amb	.34 (−)	.11	"…The kids are trying to get [um] cupcakes from the top of the fridge but her stool's falling… the kids are probably distracting the mom but I think the mom is trying to ignore it and keep doing what she's doing" (P836, PD1)
OM18	Simple declarative clause: verb phrase (Stanford POS tagger)*	Gram-const	.32 (+)	.15	"She *realized* that she *missed* the bus and she *go* back to home and *tell* his dad to *drive* her to the school" (P827, SR1)
OM19	Variation in verb ambiguity (WordNet)	Voc Spec-Sim-Amb	.30 (+)	.14	"When they *were running* Mac was *blocking* Pablo and Mac *said*, '*Get* out of the way loser,' … Ariana said, 'You can *join* us to *make* it on the track and field team,' then they *became* friends (P944, SR2)
OM20	Mean valence for verbs, from unpleasant to pleasant (Warriner)	Voc Aff-Sent	.26 (+)	.05	"…I *think* this family *is* a bit wealthier than the average family because [uh] they *get to afford* [uh] to *wear* nice clothes … It looks like they have quite a big family and *enjoys* music and they *love* [um] unhealthy food like cupcakes (P824, PD1)

Note: MAE = mean absolute error; SD = standard deviation; Dir.=direction of prediction (positive or negative predictor of reading comprehension outcome). * normalized by total number of words in each response.

Unlike the oral/regular data, where grammatical constituents were the top five predictors, vocabulary affect and sentiment was the top predictor (OM1) in the oral/modified dataset, specifically the mean arousal of nouns (as tagged in Warriner et al.'s corpus). Greater values of arousal are associated with excitement, while lower values are associated with calmness. Mean arousal of nouns is a negative predictor, so higher values (indicating excitement) predicted lower values for reading comprehension. It is possible that overly excited language relates to a loss of self-regulation when engaging in language and literacy tasks; this merits further investigation. Arousal rating of verbs (OR4) was also a top negative predictor. Mean dominance of nouns (OM7) was a top negative predictor, with higher dominance values ("in control") associated with lower reading comprehension scores, and lower dominance values ("being controlled") associated with higher reading scores. Valence, from unpleasant to pleasant, also was important in this model, with variation in noun valence (OM10) a negative predictor and mean valence for verbs (OM20) a positive predictor. The latter is slightly unexpected since the strong negative sentiment of words (OM16) was among the top 20 predictors; however, strong negative sentiment could be drawing on nouns rather than verbs.

Grammatical constituents also played an important role in this model, with three forms of verb phrases (OM2, OM3, and OM18) emerging as important positive predictors. These positive predictors were somewhat simple verb phrases; however, a verb phrase with a subordinating conjunction (OM5) was a negative predictor. This is unexpected given that subordination in general tends to be associated with greater grammatical complexity. A review of narrative samples with higher scores in this feature suggest it may be an artifact of grammatical errors, e.g., "their mother busy [uh] *cook* cooking their her children's breakfast *while* the children's what the boy is playing with the dog…" (P828, PD1).

Regarding nouns, unlike the oral/regular dataset, where a noun phrase consisting of determiner and plural noun was a positive predictor, here (OM8) it is a negative predictor. As well, noun phrases consisting of noun phrases and prepositional phrases (OM13) negatively predicted reading comprehension. These features could potentially be considered formulaic, which may be negatively associated with literacy attainment. Yet, two single or mass nouns in sequence (OM6) was a positive predictor. One other grammatical constituent, *wh*-adverb phrases (OM15) was also negatively associated with reading. As the provided sample in Table 9 suggests, the repeated use of "when" takes the role of listing events rather than constructing a coherent narrative. Two

grammatical complexity metrics were among the top predictors in this model: dependent clauses per clause (OM11) and the length of prepositional phrases (OM12), both associated with positive reading outcomes.

Vocabulary richness and range were represented by mean word frequency (OM9) and verb familiarity (OM14), which both performed as expected with higher values predicting lower reading scores. The variation in vocabulary specificity (OM17) was also a top predictor, like the oral/regular model, but here unlike there it is a negative predictor. Lastly, the variation in word ambiguity (OM19) was a positive predictor.

Table 10

Twenty most important NLP-extracted grammar and vocabulary features in the text-elicited/regular dataset, predicted through random forest regression permutation (n=99)

ID	Feature (source)	Feature type	MAE (Dir.)	SD	Example student response with high indicator for the feature
TR1	Count of verb phrases (Stanford POS tagger)*	Gram-const	.70 (−)	.25	"Bats *help* humans by *reducing* insect population, spred seeds, and *provide* medicen from there siliva." (P881, Open nonfiction prompt)
TR2	Average word meaning similarity (WordNet Lin-Brown method)	Voc Spec-Sim-Amb	.56 (+)	.11	"Bat help humans because they have cures for sickness. They help kill bugs so we don't get that many bites" (P885, Open nonfiction prompt)
TR3	Familiarity of nouns (Bristol and Gilhooly-Logie)	Voc Rich-Range	.54 (−)	.17	"I think *kids* should not be on a *phone* for a long *period* of *time*… kids are way to addicted to *games* and I understand because i have a *phone* but instead of trying to join your *friend* in front of a *screen* you could actually go outside… (P867, Writing)
TR4	Number of clauses (Lu)*	Gram-com	.54 (−)	.22	"I think that this text is not very intresting. I dont like it becaus it is not that intresting to me and there is nothing cool about this text" (P775, Nonfiction passage interest prompt)
TR5	Clause introduced by subordinating conjunction: Wh-noun phrase → simple declarative clause	Gram-const	.43 (+)	.14	"When the kids found out *what* their parents signed them up for the felt rando" emotions." (P945, Writing)
TR6	Verb phrase: past-participle verb → noun phrase	Gram-const	.31 (+)	.16	"I feel this way because she was a young women who *changed history* by swimming so far." (P915, Narrative interest prompt)

TR7	Simple declarative clause: noun phrase → verb phrase followed by a period (Stanford POS tagger)*	Gram-const	.26 (−)	.12	"I think it is good and bad to go on social media. It's good because you can contact people. But it can't help you're eyes or head" (P886, Writing)
TR8	Kurtosis of noun ambiguity (WordNet)	Voc Spec-Sim-Amb	."4 (+)	.06	"I think that social *media* is bad for young *kids* because if you are looking at a big or small *screen* for a long time it starts to hurt. Your *head* might hurt from the *brightness* because of all the cellular *waves* that are going back and forth. (P904, Writing)
TR9	Moving average type-token-ratio over a window 30 words wide*	Voc Rich-Range	.17 (+)	.06	"…social media is bad for young people because their minds have not fully devoloped and they cannot make decisions as easily…Its also good to do something you enjoy doing that is physical productive to distract yourself from checking on your device ever few minutes." (P851, Writing)
TR10	Verb ph[ra]se: non-3rd person singular present verb and simple declarative clause (Stanford POS tagger)*	Gram-const	.16 (−)	.05	"*They make* plants Grow more. *they keep* the food chain balanced" (P935, Nonfiction open prompt)
TR11	Verb phrase: base form verb → simple declarative clause (Stanford POS tagger)*	Gram-const	."5 (+)	.05	"Bats could help humans by not only reducing the mosquito population, but also helping humans with saliva which is sometimes used *to help* with ehart or blood problems. They also help because since they eat mosquitoes, people won't need *to buy* pesticides as much." (P799, Nonfiction open prompt)
TR12	Age of acquisition of nouns	Voc Rich-Range	."4 (+)	.03	"Bats help *humans* because the keep the *insect levels* down, and they spread *seeds* of *plants* around, finally the *vampire bats* help us by having a *chemical* in *saliva* that keeps *blood* going" (P905, Nonfiction open prompt)
TR13	Count of Personal pronouns (Stanford POS tagger)*	Gram-const	.14 (−)	.05	"How old were *you* when *you* wrote this story" "How did *you* know all this" "Why did *you* write a biography about *her*" (P772, Narrative questions for the author)
TR14	Variation in the ambiguity of nouns (WordNet)	Voc Spec-Sim-Amb	.14 (+)	.04	"Now, this is a *question* that *people* have always been thinking. 'Is social *media* mostly good our bad for young *people'* Well, there are 2 *sides* to every *case*. …. From communicating and live streaming *videos* on youtube to posting and looking at *pictures* on Instagram" (P750, Writing)

TR15	Not-in-dictionary words	Voc Rich-Range	.13 (−)	.05	I would *describe* mailyn's *charcter* as a *coraguse* girl because she will willing to swim from New your all the way to toronto. (P857, Narrative open response)
TR16	Strong positive sentiment of words (MPQA)	Voc Aff-Sent	.12 (+)	.04	"I found it interesting because the text use very descriptive words that made it interesting, but I prefer to read fiction adventures and mysteries over this non-fiction text." (P860, Narrative interest prompt)
TR17	T-units (Lu)	Gram-com	.11 (−)	.04	This is why bats help humans. Bats help humans for the following reasons. when they suck blood from the humans their soliva gose into the body wich is good for humans. Another reason is that they use their soliva for medicin. Those are the reass help humans" (P850, Nonfiction open prompt)
TR18	Noun phrase consisting of two noun phrases connected by a coordinating conjunction	Gram-const	.11 (−)	.04	"Having *a phone and social media* could distract you from doing *your Homework and your school work*, and it could Hurt *your eye And Brain*. another good think is you can contact with *your mom and dad* when they need you. But social media wastes time and on that time you can do way better stuff with *your Friends and family*." (P767, Writing)
TR19	Verb phrase consisting of non-3rd person singular present verb	Gram-const	.10 (+)	.03	*they help* because.1in paragraph 4 it said one of the types of bats gathers and plants seeds to help plants grow.2 in paragraph 2 it said *bats like* to eat small bugs like Mosquitos. its good that *they eat* mosquitos so they *don't bite you* and sood. (P822, Nonfiction open response)

Like the oral/regular model, in the text/regular model grammatical constituents emerged as top predictors. Unlike the oral/regular model, though, the count of verb phrases (TR1) was a negative predictor, which may relate to task effects, as described above (listing objects in the oral tasks appears to be associated with lower reading comprehension scores, while describing actions in the oral tasks was associated with higher scores). Simple declarative clauses consisting of a noun phrase and verb phrase (TR7) at the end of the sentence were negative predictors, as were non-3rd person singular present verb with simple declarative clauses (TR10). However, three verb forms were positive top predictors: past-participle verb noun phrase (TR6), base form verb simple declarative clauses (TR11), and non-3rd person singular present verb (TR19). Like

the oral/regular model, noun phrases with two nouns connected by a coordinating conjunction (TR18) was a negative predictor. Personal pronouns were also negative predictors (TR13). Finally, simple clauses introduced with wh-noun phrases (TR5, e.g., who, what, which) were positive predictors.

The only grammatical complexity features to be among the top 20 predictors were the number of clauses (TR4) and the number of t-units (TR17), both of which are normalized by the number of words in the narrative. These represent briefness of T-units and clauses (lack of grammatical complexity) and both were negative predictors of reading comprehension.

Four vocabulary richness and range variables were top predictors, including noun familiarity (TR3), age of acquisition of nouns (TR12), not-in-dictionary words (TR15), and the moving type-token average with a 30-word window (TR9). These all behaved as expected, with greater sophistication and greater variation correlated with greater reading scores, and not-in-dictionary words showing negative association (they are likely spelling errors). Somewhat unexpectedly, the average similarity of words' meanings (TR2) was a top positive predictor, perhaps indicating cohesion in writing. Kurtosis of word ambiguity (TR8) and variation in noun ambiguity (TR14) also positively predicted reading outcomes, which is somewhat difficult to interpret but may relate to the use of some ambiguous and some non-ambiguous words. Finally, the affect and sentiment category is represented by strong positive sentiment of words (TR16), which makes sense given the tasks, which pertained to the pros and cons of social media, a young woman's successful swim across Lake Ontario, and interesting facts about bats, with a focus on how they can help humans.

Table 11

Twenty most important NLP-extracted grammar and vocabulary features in the text-elicited/modified dataset, predicted through support vector regression permutation (n=67)

ID	Feature (source)	Feature type	MAE (Dir.)	SD	Example student response with high indicator for the feature
TM1	Average word meaning similarity (WordNet LC method)	Voc Spec-Sim-Amb	.95 (−)	.38	"Social media is not good for young people because its not good to stay inside and not to exiersize outside. Also it can be bad for their eyes and if you stay inside and watch movies and play video games is bad for their eyes…." (P907, Writing)

TM2	Variation in the similarity of meaning between words (WordNet)	Voc Spec-Sim-Amb	.58 (−)	.37	“The good things about it are when you are bored you can go on social media on your device. The bad things about it are... You can get addicted to the device/electronic. Also there are alot inappropriate things on social media.” (P771, Writing)
TM3	Word imageability (Bristol and Gilhooly-Logie)	Voc Rich-Range	.46 (−)	.12	“i am the youngest in my family and i wanted a sister or a brother and i ended up with 1 brother and 3 sister” (P783, Interest prompt for narrative passage)
TM4	Weak-positive word sentiment (MPQA)	Voc Aff-Sent	.39 (−)	.17	"it's not Really good for you Because it keeps you away from you family and friends you Don’t get to sit With them and talk. And it's good Because you conect With Friends and Help them and easy to Find them...” (P767, Writing)
TM5	Average length of words	Voc Rich-Range	.31 (+)	.10	“Jada felt brave because she felt she had something to be responsible for” (P873, Open response prompt for narrative passage)
TM6	Neutral word sentiment (Stanford Sentiment)	Voc Aff-Sent	.25 (+)	.09	"I feel that this story is good education for people who's going to get a newborn baby. Also, this way, people would be able to follow these stuff?" (P949, Interest prompt for narrative passage)
TM7	Average word meaning similarity (WordNet SemCor method)	Voc Spec-Sim-Amb	.22 (−)	.10	"… If your with your family for 2 hours we should be with your family 4 hours….I watch social media 1 hour…I am going to reduce social media for only 20 min because I Love watching social media." (P924, Writing)
TM8	Verb phrase: 3rd person singular present verb → noun phrase (Stanford POS)*	Gram-const	.22 (+)	.11	“*Is a eagle owl* more powerful than a owl? *Do eagle owls* bite humans! *Are eagle owls* friends with owls?” (P948, Questions for the author, nonfiction passage)
TM9	Coordinate phrases per clause (Lu)*	Gram-com	.22 (−)	.08	“I think social media is good *and* helps kids play games *and* socialiez with peaple far away *or* not” (P864, Writing)
TM10	Number of prepositional phrases (Fraser, 2016*	Gram-const	.21 (+)	.08	“…Social media is good and bad *in different ways*…it will help *with school work* talking to people when there bored or need help. And the bad it can make people detach *from family sports* and its not very healthy to stare *at a light* to long.” (P882, Writing)
TM11	Count of coordinates (Stanford POS)*	Gram-const	.21 (−)	.08	"…people do not like it *and* they will no be their friend. They could get hacked *and* it is very bad because at some games it needs to use login information *and* people might see it.” (P790, Writing)

TM12	Variation in the ambiguity of nouns (WordNet)	Voc Spec-Sim-Amb	.21 (−)	.05	“Why is there no conflect and most stories have conflict. Do you want to make more storys like this.” (P892, Questions for the author, narrative passage)
TM13	Average ambiguity of nouns (WordNet)	Voc Spec-Sim-Amb	.20 (+)	.08	"Social *Media* is bad for young *people* because *it* will damage your *eyesight*, *you* won't spend *time* with your *family*, *you* won't have *creativity*, *you* will get addicted to your *device*, *you* start to not care about *anything*, *you* can't think well." (P919, Writing)
TM14	Verb phrase: gerund/present participle → noun phrase	Gram-const	.16 (+)	.04	“Because when I was *becoming an big sister* the same thing happend instrad I was with my grandma and grandpa” (P810, Interest prompt for narrative passage)
TM15	Verb imageability (Bristol and Gilhooly-Logie)	Voc Rich-Range	.15 (−)	.06	"… [they can know] how old you *are* wend you *move* what ages you *have* what your cars lisenplat. Thay can *rob* your hous. You will *think* no bout will do that but there *is* alot in the world." (P761, Writing)
TM16	Simple declarative: noun phrase → verb phrase (Stanford POS)*	Gram-const	.15 (−)	.11	“It's good because If *you had* an enmergency event that *you need* to talk to your friends, if *you mail* them, It's too late, but if *you e*-mail them…*It's* never late. *It is* bad because *it created* syber bulling, and *it's* harmful for eyes….” (P800, Writing)
TM17	Mean arousal of all words, from calm to excited (Warriner)	Voc Aff-Sent	.15 (+)	.08	“Kids spend alot of time playing video games online that evolves violence. If kids play games that have too much violence, stop them! When kids play games that have too much violent content in it, the kid is going change slowly.” (P950, Writing)
TM18	Average maximum depth from each verb to root (WordNet)	Voc Spec-Sim-Amb	.15 (+)	.08	“I am not *interested* in non-fiction but it's so cool that they *swallow* animals whole.’” (P760, Nonfiction interest prompt)
TM19	Count of personal pronouns (Stanford POS)*	Gram-const	.14 (−)	.04	*“*1) *You* will never meat new people/or friends. If *you* do *you* will never know their personality. 2) *You* might get angry at your parents when they tell *you* to not use *you* phone. 3) *You* will spend more time with family…” (P951, Writing)
TM20	Variation in word similarity (Wordnet)	Voc Spec-Sim-Amb	.14 (−)	.10	“Everyone knows that older brother/sister needs to protect their younger brother/sister. That's what happen if you have a younger/sister and if you care for them.” (P793, Open narrative prompt)

Note: MAE = mean absolute error; SD = standard deviation; Dir.=direction of prediction (positive or negative predictor of reading comprehension outcome). * normalized by total number of words in each response.

Unlike text/regular (where it was a top positive predictor), average word meaning similarity (TM1, TM7) is a top negative predictor in the text/modified dataset, along with variation in word meaning similarity (TM2, TM20). The average noun ambiguity (TM13) was a positive predictor. Yet, variation in noun ambiguity (TM12) was a negative predictor. Easier to interpret is average word specificity (TM18), which was a positive predictor.

Grammatical constituents do not emerge as top predictors in this model until TM8, where a verb phrase consisting of a 3rd person singular present verb followed by a noun was a positive predictor, as well as verb phrases consisting of a gerund/present participle and a noun phrase (TM14). Yet simple declarative phrases composed of noun phrases and verb phrases (TM16) were negative predictors. Two basic grammatical constituents negatively associated with reading outcomes were the count of coordinates (TM11), which was also present in the grammatical complexity category with coordinate phrases per clause (TM9). A similar result was found in the oral/regular and text/regular, which suggests again that extensive use coordinates is consistently a negative predictor of reading comprehension. In addition, the count of personal pronouns (TM19) was negative, similar to text/regular. Prepositional phrases (TM10) were a strong positive predictor.

Three vocabulary richness and range metrics were important predictors in this model, all functioning as expected: word imageability (TM3), average length of words (TM5), and verb imageability (TM15). Finally, three affect and sentiment features were among the top 20 predictors: weak-positive sentiment (TM4), which was negatively associated with reading outcomes, neutral sentiment (TM6), which had a positive association, and mean arousal (TM17), with words associated with excitement correlating with higher reading outcomes.

The top 20 predictive features for each model are discussed below, organized by feature type.

4.1.4 RQ1 Discussion

4.1.4.1 Grammatical constituents and complexity

Grammatical constituents consist of counts of certain grammatical features (and combinations of those features) normalized by the number of words in the sample. Grammatical complexity concerns the length of phrases, clauses, T-units, and sentences. Noun forms, when present in the top 20 predictors, were consistently negative predictors (9 of 11), with a few exceptions. This

suggests that despite nominalization being a key element of academic language (Fang, Schleppegrell, & Cox, 2006; Halliday, 1993), in these language samples, the use of nouns is associated with lower reading outcomes. This recalls Crossley et al.'s (2014) finding that noun use is not a focus for human raters as they evaluate writing. As a corollary, the presence of verb features in the top 20 predictors tended to be positive (11 of 15). That most positive verb predictors were in the oral datasets could also be related to task effects, because the oral tasks ask participants to provide narratives that require action words (although some participants instead used listing or formulaic language), while the text-based tasks required persuasive and descriptive language (i.e., not focused on actions per se, but instead focused on character traits or global inferences).

Children learn nouns before verbs (Gentner, 2006). As Gentner (1982) argues, nouns "have a particularly transparent semantic mapping to the perceptual-conceptual world," while verbs "have a less transparent relation to the perceptual world" (p. 328). In both oral models, the count of simple declarative clauses that consist of a verb phrase was a positive predictor. While this is a simple grammatical constituent, in the oral tasks specifically, the use of verbs appears to represent an understanding of the actions and interactions that took place in the pictures participants described and the stories they comprehended and retold. However, for writing tasks, predominance of verbs and verb phrases did not appear to play an important positive role (although certain characteristics of verb phrases did). In fact, count of verbs was the top negative predictor for text/regular. Again, this may relate to the nature of the tasks, where action verbs were not required.

While the COVFEFE pipeline outputs extremely fine-grained information about grammatical constituents, such as whether a verb is 3rd-person or not, and what parts of speech follow it, there were no clearly discernable patterns in terms of whether 3rd-person or non-3rd person usage, or noun or verb phrases that were more simple or more complex (having additional embedded phrases) are with stronger reading comprehension. Yet the grammatical constituents did emerge as top predictors in all models, especially those using the regular datasets. Thus, it appears that the only potentially generalizable aspect of the noun and verb elements is that grammatical elements with verbs tended to predict positively, while elements with nouns tended to predict negatively, but when fine-grained noun and verb characteristics emerge as top predictors, they may be task-specific and not (as) generalizable.

Also of interest is that in all models except oral/modified, grammatical constituents containing the simple combination of noun phrase followed by a verb phrase was a negative predictor (in oral/regular, it was a noun phrase consisting of a noun phrase and verb phrase; in both text models, it was a simple declarative clause consisting of a noun phrase followed by a verb phrase). These two forms share similarities but are not identical. An example of the former, a noun phrase consisting of a noun phrase and verb phrase, from PD1: "… I see *a dog giving* a boy a ball a girl in this picture I see *a girl tryin'* to get cupcakes in this picture there's *a mom making* food" (P921). Although the clauses may be lengthy, the language becomes formulaic as the same grammatical form is repeated. The latter form, simple declarative clauses consisting of a noun phrase followed by a verb phrase, is a very simple grammatical structure that contains just that. Because these counts are normalized by the number of words, and because, in general, complex clauses are positively correlated with the reading outcome, it is intuitive to interpret that these simple grammatical structures are associated with lower reading comprehension outcomes.

Personal pronoun counts were negative predictors in both text models but not the oral models. In oral language, referring to people, objects, or concepts by pronouns is common, while written language often requires more specific use of named people, objects, or concepts. Crossley et al. (2014) analyzed pronouns as anaphoric referents, finding they are positively correlated with writing coherence. Grant and Ginther (2000) examined pronoun type (first-person, second-person, third-person) in writing by adults learning EAL and found higher-level students used fewer second-person pronouns ("you") and more third-person pronouns ("he" or "she"). The present study does not track anaphoric referents or types of personal pronouns, but the results suggest that overuse of pronouns in writing (but not in speaking) is associated with lower reading outcomes, and that this may relate to a lack of specificity that is necessary in quality writing.

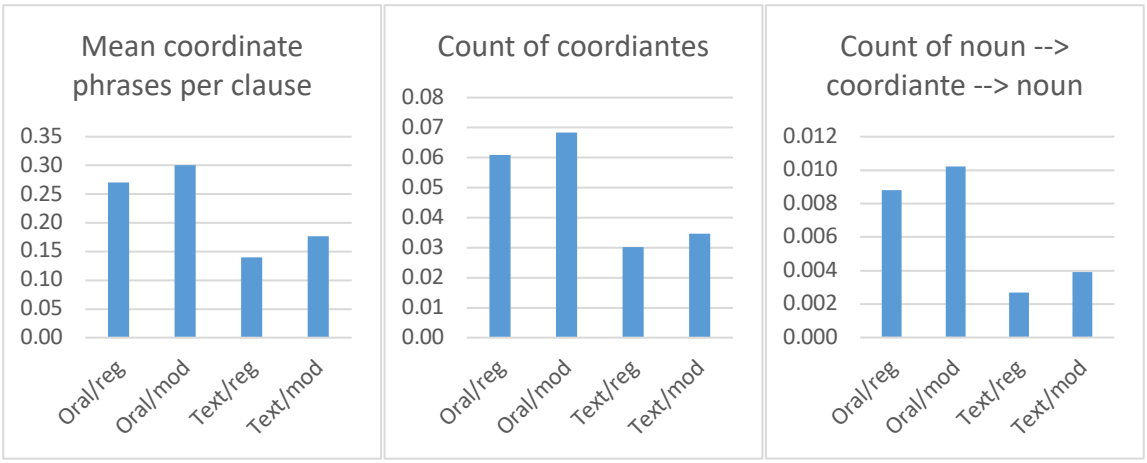

Figure 18. Mean use of grammatical coordinates across the four models (counts are normalized)

Coordinating conjunctions, and phrases with that part of speech, were consistently negative predictors in all datasets except oral/modified. This could relate to a tendency for "listing" concepts using the word "and", which may not require deeper cognitive processing like formation of dependent clauses. Yet, as seen in Figure 18, the oral/modified data had more instances of coordinates (and coordinate-containing phrases) than any other dataset. Even though coordinating conjunction features were ubiquitous in oral/modified data, they were not found to be important predictors of reading comprehension in that dataset. Because the current analysis uses data on a 2 x 2 matrix (oral/text by regular/modified), it is not possible to determine whether the difference for the oral/modified dataset is due to differences in written and spoken language, or due to changes in language and cognitive development (participants who completed BALA modified were younger or were identified as benefiting from a modified reading comprehension assessment by their teachers, due to reading, language, or cognitive factors). Nonetheless, this question would benefit from further research.

A moderately discernible pattern is that in the modified datasets, the number (for text-elicited) and length (for oral-elicited) of prepositional phrases were both positive predictors, aligning with Crossley et al. (2016). For oral/regular, a prepositional phrase with a single preposition (presumably relatively simple in terms of grammar) was a negative predictor, while more complex grammar — a verb phrase consisting of a gerund/present participle verb, then noun phrase, and then prepositional phrase — was a positive predictor. Interestingly, the normalized count of prepositions was relatively similar in all four datasets, between .08 and .10, but generally speaking, prepositions were a positive predictor in the modified data but not the regular

data (except for complex forms). Like the discussion of coordinate conjunctions above, the differences in predictive relationship between prepositional phrase use and reading comprehension may relate to language development or maturity.

Regarding grammatical complexity, for the text/regular model, the number of clauses in the entire sample, and the number of T-units, normalized by the length of the total sample were important negative predictors. These features are counts normalized by the length of the sample. A greater number of T-units and clauses in a sample means that the length of each clause/T-unit is short; shorter clauses and T-units are less grammatically complex. Similarly, for the oral/modified model, the number of dependent clauses per clause was a positive predictor.

In sum, the grammatical features relating to coordinating versus complex clauses, and longer versus shorter clauses and T-units, are intuitive to interpret: from the given literature their relationship with reading is fairly straightforward. However, the other results discussed here – features relating to nouns tending to be negative predictors, while features pertaining to verbs tending to be positive predictors, pronouns as negative predictors in the text models, and the mixed results for prepositional phrases – require further research to understand if indeed these findings are generalizable.

4.1.4.2 Characteristics of vocabulary

The top vocabulary richness and range features all performed as expected across the four models. Positive predictors included the Honoré statistic (oral/regular), the 30-word moving average of the type-token ratio and age of acquisition of nouns (text/regular), and word length (text/modified). The top negative predictors were imageability of all words (oral/regular and text/modified), verb frequency (oral/regular), word frequency and verb familiarity (oral/modified), not-in-dictionary words and familiarity of nouns (text/regular), and verb imageability (text/modified). In terms of specific parts of speech, vocabulary range metrics (imageability, familiarity, or frequency) relating specifically to verbs were top predictors in all models except text/regular, where noun familiarity and age-of-acquisition were found to be more important. That age-of-acquisition is a subjective rating based on adults' memory of when they learned a word could dilute the predictive value for the modified data (i.e., for less sophisticated words that are learned at a younger age, the adults' memory of the exact age may not be as

accurate); however, the text-regular data contained the most "adult-like" language, so the presence of age-of-acquisition as a top predictor in that model is not unexpected.

Fifteen vocabulary specificity, similarity, and ambiguity features were present in the top predictors across models. Five of the 15 were averages, and 9 of 15 were standard deviations, which represent variation in that feature. One top predictor was a kurtosis function, which represents a peakedness across all the words in that dataset's samples. Table 12 summarizes the commonalities and differences in this category of features across the four models.

Table 12

Top vocabulary specificity, similarity, and ambiguity features across four models

Model	Specificity	Similarity	Ambiguity
Oral/regular	Variation in maximum depth from word to hypernym root (+) Variation in minimum depth from word to hypernym root (+)	Variation in word similarity (+)	
Oral/modified	Variation in max depth from word to hypernym root (−)		Variation in verb ambiguity (+)
Text/regular		Average word similarity (+)	Variation in noun ambiguity (+) Kurtosis of noun ambiguity (+)
Text/modified	Average maximum depth from verb to hypernym root (+)	Average word similarity, two methods (−) Variation in word similarity, two methods (−)	Average noun ambiguity (+) Variation in noun ambiguity (−)

The averaging functions offer the most straightforward interpretation of these features. The average maximum depth from word to root hypernym, which measures semantic specificity, is a positive predictor for the text/modified model. This is expected, as this is a measure of word specificity; although, as discussed earlier, WordNet specificity is not always correlated with

vocabulary sophistication. Average noun ambiguity was a positive predictor in the text/modified model. This aligns with the literature discussed above on the developmental trajectory for increased use of ambiguous words (Casas et al., 2018). It is important to note that Casas et al. found no significant differences after about 5 years of age; therefore, it is sensible that this feature was a positive predictor for the modified dataset, which contains information about younger and/or less linguistically developed participants. The contrasting results for average word similarity (positive predictor in text/regular but negative in text/modified) are difficult to interpret, but may relate to the effects that the use of similar words produces. Referring to the participants' language samples provides some insight. Both samples presented below are taken from the social media writing task and both scored high for average word similarity. The first sample is from the modified administration (wherein average word similarity negatively predicted reading comprehension), and the second from the regular administration (average word similarly positively predicted reading comprehension).

Sample 1: P843, Modified administration

A lot of social media is bad bad for young people. That is even though it is called social midia you can't be really social with it. You can only text it. It is better when you meet the person for your self. I have never used social media but I have used computers. When I use a lot of it I get very angry. I start to feel confused and kind of sick. I get a lot of mixed feelings. I start to feel tired. When I use a lot of computers, I don't get time to spend with my family. But when I don't use computers & feel feel happy…

Sample 2: P938, Regular administration

Personally I believe that social media is mostly not good for children and teens but i still support people who say that it helps them to calm down since many people have different opinions about different things. Here are some examples on this opinion: 1. Social media can have many different effects on people, one of them being stressing them out. Social media can be stressful because often times if you use and post on it oftaine you might feel inclined to post even when you are in a situation in which you feel uncomfortable about doing so…[participant provides 2 more examples]

While both samples repeat certain words, the regular administration sample appears to use repetition to build cohesion, while the modified administration sample uses the same words repeatedly without the positive effect of cohesion, and also literally repeats words (e.g., "bad bad" and "feel feel"). From this example it is possible to see how word similarity can be positive in some cases and negative in others.

The variation and kurtosis functions are also difficult to interpret, but the they appear to be consistently positive in the regular models but mostly negative (except for one such feature) in the modified models. This could again relate to language development ,or maturity, where, for participants who completed the regular BALA reading comprehension measure, more variation in specificity, ambiguity, and similarity means better control over these constructs; but for those who completed the modified version, the variation is perhaps more random and not as well controlled.

The vocabulary affect and sentiment features are based on corpora developed either through ML (e.g., learning the sentiment of movie review language by using the numeric review [number of stars] as the label) or via surveys. Table 13 summarizes the affect and sentiment features across the four models.

Table 13

Vocabulary affect and sentiment across four models

Model	Valence (unpleasant to pleasant)	Arousal (calm to exciting)	Dominance (controlled to in control)	Sentiment
Oral/regular				Strong negative (two methods) (+)
Oral/modified	Mean of verbs' valence (+) Variance of nouns' valence (−)	Mean of nouns' arousal (−) Mean of verbs' arousal (−)	Mean of nouns' dominance (−)	Strong negative affect (+)
Text/regular				Strong positive (+)

Text/modified	Mean arousal (+)	Weak positive sentiment (−) Neutral sentiment (+)

Broadly speaking, these features were more predictive for the modified models than for the regular models, with the only such features emerging as top predictors for the regular models being positive/negative sentiment. Noting that negative sentiment is a positive predictor in both oral datasets, and positive sentiment being a positive predictor in text/regular – suggests the nature of the task may influence which sentiment is associated with higher reading achievement. This sample (P860) from the text/regular data had a high score for strong positive sentiment:

> "I would say Marilyn is a determined person. Even when she was exhausted she decided to keep on going. Also Marilyn after the 2 hours and 57 minutes of swimming across Lake Ontario Marilyn was probably extremely happy that she was on shore." [Writing]
>
> "Why do YOU find Marilyn Bell's swim inspiring?" [Questions for the author]

While this sample (P928) from oral/regular had strong negative sentiment:

> Ariana and Pablo were practicing for the track and field tryouts the next week and Pablo started running when the school bully stood in front of him and he was too scared to talk to the bully [um] then Ariana came and she helped Pablo tell off the bully [um] and then Pablo asked if mac the bully wanted to race with them …

As mentioned earlier, a writing prompt that refers to an inspirational swim across Lake Ontario is likely to elicit positive sentiment, while writing about a school bully is likely to elicit negative sentiment. While these are strong predictors for these particular oral- and text-based tasks, sentiment analysis as a predictor of reading comprehension skill does not appear to be generalizable. However, the level of sentiment used (regardless of valence) would be an interesting avenue for further research.

Positive valence (which ranges from unpleasant to pleasant) of verbs was a positive predictor for oral/modified, but mean arousal (nouns and verbs) and dominance (nouns) were negative predictors. To illustrate, following are two samples from SR2 from the oral/modified data, the

first with a high mean verb arousal score, and the second with the lowest mean verb arousal score of all participants in that dataset:

> Ariana and Pablo *was racing* each other for the for *being chosen* for the track and field team Pablo was *running* but the mac mac *stopped* his way then mac *was* the bullying person he *said get out of the way* loser and Ariana said he he *is* not a loser and if you *want* if you if you *want* to *race* you could and mac *raced* each other for the end of the recess… (P828)

> Pablo *was* too afraid to *speak* up for himself but but then but then Ariana *was* instead of being a bystander *stood* up for him and and and made Mac stop *picking* on Pablo after that Pablo *felt* stronger and more brave so he *asked* the bully do you want to *practice* with us so the bully *answered* that he he *had been wanting* to be on the track field for very long and then all three of them *started practicing* and after all the practice and hard work all three of them *made* it onto the school track and field team (P806)

High verb arousal (as seen in the first sample) tended to be associated with lower reading comprehension scores, and vice versa. With regard to self-regulation, it is possible that content of the oral tasks (bullying, being late, and somewhat chaotic scenes in the kitchen and on the street) was exciting and that students who were able to maintain a more calm, even abstract, tone, are able to process the language efficiently, which could relate to reading comprehension success. However, like the word similarity, specificity, and ambiguity results above, the specific results for affect and sentiment are not conducive to straightforward interpretation. Again, these features emerged as more important in the modified datasets (especially the oral/modified) than regular datasets, suggesting that emotions and affect in language are playing a more important role for learners who are younger or less linguistically mature.

In summary, there were three types of lexical features in these analyses: vocabulary richness and range, vocabulary specificity, similarity, and ambiguity, and vocabulary sentiment and affect. The richness and range metrics predicted reading comprehension in alignment with the published research in this area: use of vocabulary of higher familiarity, frequency, and imageability was associated with lower reading proficiency, while use of longer words, and words with higher age-of-acquisition had the opposite relation. Honoré's statistic, which is a function of how many words were used only one time in each sample, also was a positive predictor. Interestingly, the

only other important lexical diversity feature among the four models, a moving type-token average with a 30-word window, emerged as a top predictor only in text/regular. Revisiting the correlations in Table 6, all the moving type-token ratios were positive correlates of reading comprehension in text/regular, while the relationship in other three datasets was weaker or even negative. The text/regular can be assumed to represent the older participants, and/or those whose teachers did not recommend them for a modified version of the reading assessment. And, production of written language tends to be more abstract and dense than oral language. Therefore, that moving type-token average is a predictor here and not in other datasets is not necessarily unexpected, since these participants are on the whole more advanced, drawing from a greater range of vocabulary, and as a text-based dataset they are operating in a more abstract mode than when speaking.

As a corollary, important predictors of reading comprehension in the modified datasets were more likely to be affect or sentiment than vocabulary richness and range. It is possible that self-regulation plays a role in the types of sentiment that are associated with positive reading outcomes. However, the predictive power of the sentiment and affect variables included here appear to be local associations, specific to the tasks, rather than a generalizable relationship. Further research is needed to ascertain the role that sentiment and affect play. Similarly, while vocabulary specificity and use of ambiguous words as evaluated by WordNet appear to play a positive role, in accordance with the literature, future investigations focusing on this question could better understand the role these and word similarity play, as well as the variation of each, in contributing to successful language and literacy development.

4.2 RQ2: interactions between top predictors and demographic factors

The second set of research questions asks: Is there significant interaction between the top lexical and syntactic features and students' length of residence in Canada, or their multilingual proficiency in languages other than English, in predicting reading comprehension? If interactions exist, how do they differ across the four models?

The top five predictors in the oral/regular model were permuted individually, and then the two demographic variables under consideration (multilingual proficiency and years in Canada) were also permutated individually. Rows and columns labeled "Individual permutation" in Table 14

provide this information. Specifically, the individual permutation refers to increase in mean absolute error when that variable is randomly permutated. Then, the pairwise permutations (with each top predictor and each demographic variable) are examined to determine if the mean absolute error decrease is greater for the two variables together than for each individually.

The individual permutation values (i.e., decrease in mean absolute accuracy) in this section are not identical to the models described in the first research question because these were performed in the Ranger package in R, rather than using Scikit-learn in Python. The Ranger package exclusively performs random forest. Although the results for the first set of research questions suggest that two of the datasets were best analyzed by non-random forest models, Ranger was selected for this research question because it has the capability to perform pairwise permutation, whereas Scikit-learn does not. This decision was also made because this investigation into pairwise permutation is exploratory only, in that it looks for relative increases in MAE compared to the permutation of individual variables. Its role was to identify interaction candidate variables to be checked post-hoc with a confirmatory multiple regression approach. (This is necessary because due to the sample size, all interactions could not be tested using multiple regression.) The top five features selected from each model were those from the original model development (first set of research questions). Ranger's default model-building settings were used.

Table 14

Pairwise permutation (interaction) results for the oral/regular model

	Individual permutation	Multilingual proficiency	Years in Canada
Individual permutation	--	.03	.13
Content density: proportion of nouns, verbs, adjectives, and adverbs	.46	.34	.33
Prepositional phrase consisting of a preposition or subordinating conjunction	.21	**.34**	**.36**
Noun phrase: noun phrase → coordinating conjunction → noun phrase	1.23	**1.47**	1.09
Noun phrase: determiner → plu[ra]l noun	.13	.12	.08
Verb phrase: non-3rd person singular present verb	.62	**.83**	**.83**

Note: Bolded figures represent an increase in mean absolute error for pairwise permutation that is greater than either of the individually permuted variables.

As visible in Table 14, five interactions (bolded) in the oral/regular dataset were identified as having greater decreased MAE when permutated in tandem than individually. Multiple regression analysis tested whether significant interactions existed between bolded combinations in Table 14, (adjusted R^2=.15, $F(10, 84)$=2.65, p<.01). No interactions were statistically significant, but two individual features were found to negatively predict reading comprehension: the use of noun phrases consisting of a noun phrase, coordinating conjunction, and noun phrase (β = -487.60, p<.01) and the use of prepositional phrases consisting of a preposition or subordinating conjunction (β = -3681.13, p<.05).

Although not significant, a visual of the interaction of the latter with the dichotomized Years in Canada variable is shown for demonstration in Figure 19. The pink trend line represents participants who have lived in Canada for ten or more years, while the blue trend line represents participants who have lived in Canada for less than ten years. Both groups show a negative relationship between use of prepositional phrases consisting of prepositions or subordinate conjunctions, and reading comprehension. The group living in Canada for less than 10 years appears to have a slightly steeper negative slope, indicating a slightly stronger negative correlation, but again, this difference was not statistically significant.

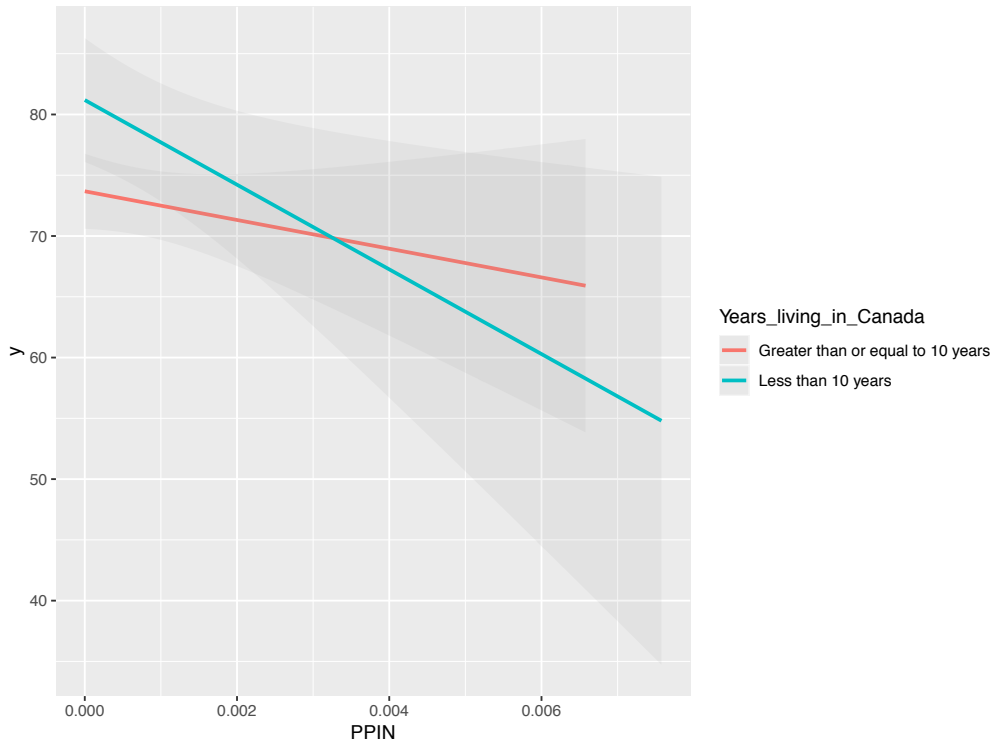

Figure 19. Relationship between reading comprehension scores (y-axis) and count of prepositional phrases with prepositions or subordinate conjunctions (x-axis) for oral/regular dataset, by years living in Canada

Table 15

Pairwise permutation (interaction) results for the oral/modified model

	Individual permutation	Multilingual proficiency	Years in Canada
Individual permutation		.01	.09
Verb phrase: "to" → verb p^{hr}ase	1.00	**1.32**	**1.22**
Verb phrase: non-3rd person singular present verb → subordinating conjuncti^{on}/clause	1.71	**1.89**	**1.88**
Verb phrase: 3rd person singular present verb → verb phrase	.38	**.39**	**.65**
Mean arousal of nouns, from calm to excited	2.24	**2.54**	**2.41**
Mean arousal of verbs, from calm to excited	.22	**.49**	**.68**

Note: Bolded figures represent an increase in mean absolute error for pairwise permutation that is greater than either of the individually permuted variables.

Multiple regression analysis on the oral/modified dataset tested whether significant interactions existed between the ten bolded combinations in Table 15, (adjusted R^2=.25, F(17, 52)=2.38, p<.01). Mean arousal of nouns, from calm to excited (β = -48.73, p<.05) was a significant negative predictor, as well as the interaction between the use of verb phrases consisting of "to" and another verb phrase, and years in Canada (β = 1470.16, p<.02). Figure 20 represents this significant interaction. For participants living in Canada for 10 or more years, this feature had a slightly negative association with reading comprehension. For participants who lived in Canada for 10 or more years, the slope was visibly positive.

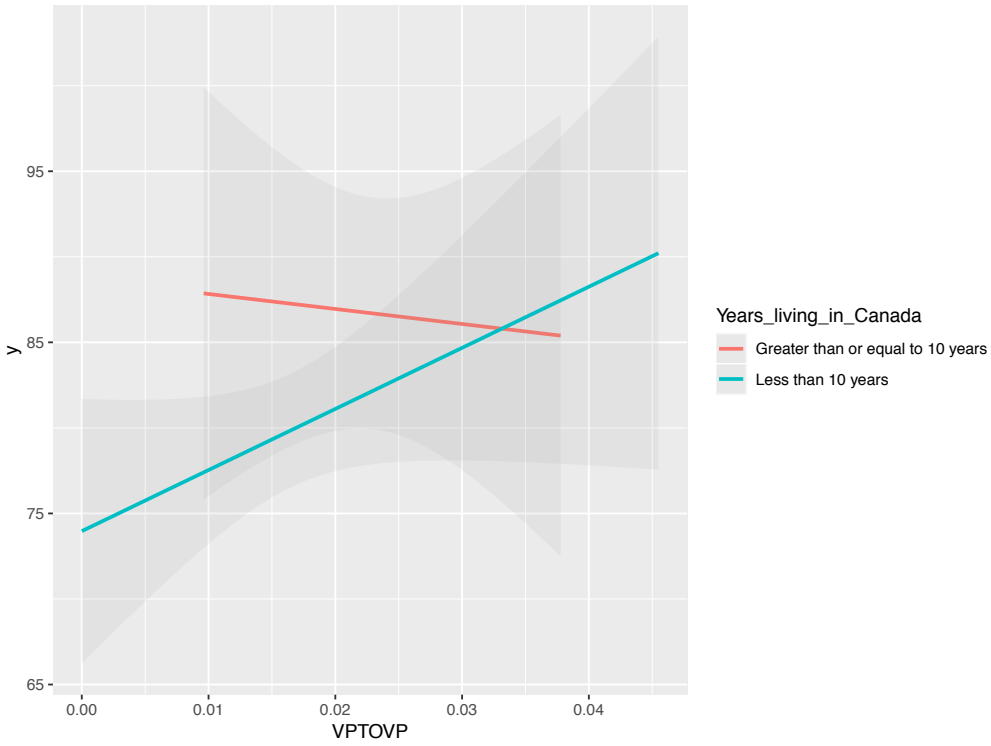

Figure 20. Relationship between reading comprehension scores (y-axis) and count of verb phrases containing the word "to" (x-axis) for oral/modified dataset, by years living in Canada

Table 16

Pairwise permutation (interaction) results for the text/regular model

	Individual permutation	Multilingual proficiency	Years in Canada
Individual permutation	--	-.02	.04
Average word meaning similarity	.72	.53	.55
Number of clauses	1.78	**1.82**	1.73

Familiarity of nouns	2.25	**2.44**	**2.28**
Count of verb phrases	2.18	2.16	**2.55**
Clause introduced by subordinating conjunction: Wh-noun phrase simple declarative clause	1.81	1.46	1.67

Note: Bolded figures represent an increase in mean absolute error for pairwise permutation that is greater than either of the individually permuted variables.

Multiple regression analysis on the text/regular dataset tested whether significant interactions existed between the four bolded combinations in Table 16, (adjusted R^2=.21, $F(9, 88=9)$=3.84, p<.001). Although 21% of the variance in reading comprehension scores were explained by the three linguistic features (number of clauses, noun familiarity, and count of verb phrases) and two demographic variables, none of the individual variables nor their interactions were statistically significant. Noun familiarity approached significance (β = -.38, p=.07), and its interaction with total multilingual proficiency was not significant but approached significance (β = .37, p<.10); for visual reference this is plotted in Figure 21. Participants who reported total language proficiency of 7 or greater (the self-assessment options ranged from 1 to 5 for each language) are represented by the green line, and participants who reported a 6 or less on for total multilingual proficiency are represented by the pink like. Noun familiarity has a strong negative predictive value for the group with less multilingual proficiency, while it has a slight positive trend for students with greater multilingual proficiency. The expected direction of prediction is negative, with increased familiarity being associated with lower scores on the outcome variable. If multilingual students do not show the same relationship, is it possible that they are drawing on their multilingual proficiencies when producing language, but these are not translating into reading skills? In addition, the distribution of noun familiarity for the group with greater multilingual proficiency is truncated versus the wider range for the less multilingual group. It remains to further research to understand why this truncation may exist, especially considering that the greater multilingual proficiency group's distribution falls near the center of the less multilingual proficiency group. Again, this interaction was not statistically significant; however,

the trend merits investigation because it is unexpected that differences between noun familiarity and reading comprehension would exist between these two groups.

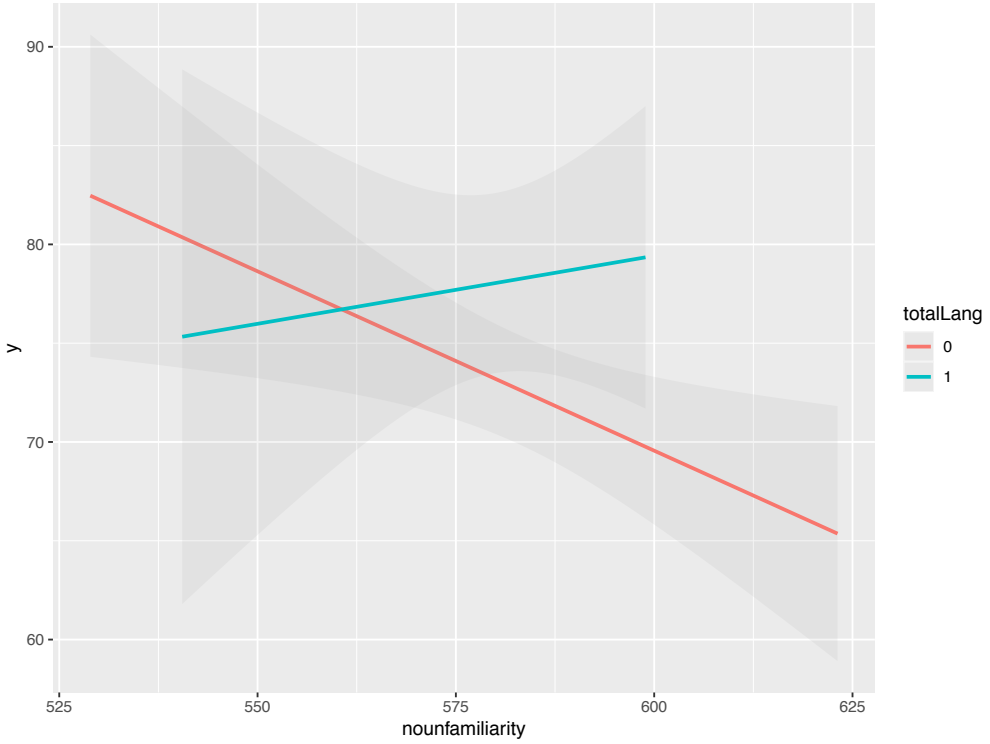

Figure 21. Relationship between reading comprehension scores (y-axis) and average familiarity of nouns (x-axis) for text/regular dataset, by total multilingual proficiency

Table 17

Pairwise permutation (interaction) results for the text/modified model

	Individual permutation	Multilingual proficiency	Years in Canada
Individual permutation	--	-.01	.07
Average word meaning similarity	3.38	**3.44**	**3.41**
Word imageability	.82	.60	**.84**
Weak-positive word sentiment	.95	.71	**1.13**
Variation in the similarity of meaning between words	2.22	**2.44**	**2.67**
Average length of words	1.01	.47	.77

Note: Bolded figures represent an increase in mean absolute error for pairwise permutation that is greater than either of the individually permuted variables.

Multiple regression analysis on the text/modified dataset tested whether significant interactions existed between the six bolded combinations in Table 17, (adjusted R^2=.21, F(10, 56)=2.78, p<.01). The word imageability feature was significant (β = -.11, p<.01) as well as the interaction between that feature and years in Canada (β = .17, p<.05). This interaction is depicted in Figure 22. For participants living in Canada for less than 10 years, word imageability is a strong negative predictor of reading comprehension. For participants living in Canada for 10 or more years, there is a very slight positive trend in the association between imageability and reading comprehension.

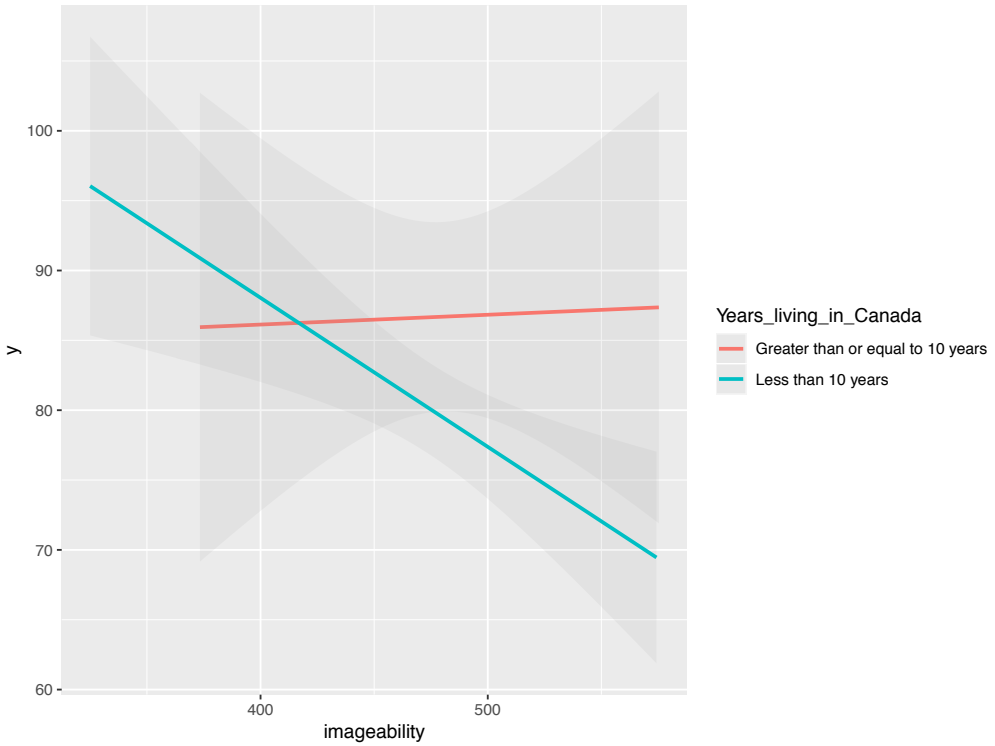

Figure 22. Relationship between reading comprehension scores (y-axis) and average word imageability (x-axis) for text/modified dataset, by years living in Canada

4.2.1.1 RQ2 Discussion

This research questions asked whether significant interactions were present between two demographic variables and top lexical and syntactic features in predicting reading comprehension outcomes. In both of the modified models, one interaction between a language feature and a demographic variable was significant. For oral/modified, the use of a complex verb phrase significantly interacted with a dichotomized Years in Canada variable. The group that had

been in Canada for 10 years or more showed a slightly negative slope (which was also truncated), while the group that had been in Canada for less than 10 years had a strong, positive slope. For text/modified, the significant interaction was between word imageability and Years in Canada. (The multilingual proficiency variable did not have a significant interaction with any of the tested linguistic features, but it did approach significance in an interaction with noun familiarity.)

While these are only two significant interactions, it is interesting to note that the group that had been in Canada longer had flatter slopes for both interaction models. It is left to future research to examine whether this difference based on immigration background is a consistent finding with other features or in other datasets. Additionally, of the two linguistic variables had significant interactions with the demographic variables, one was grammatical and one was based on vocabulary. It would be fruitful to examine these further to determine if patterns exist with regard to differential functioning of syntactic or lexical variables across other demographic groups. That both of the interactions occurred with the modified datasets is also of interest.

Differential response functioning as a field of research and practice examines how assessments function differently for different subpopulations. Little to no research has been performed on the use of fine-scaled linguistic features to predict literacy outcomes, although Zhang, Dorans, and Rupp have published research (2017) examining the functioning of an automated essay scoring algorithm known as e-rater. Because NLP applications are becoming increasingly common in both high-stakes and low-stakes educational assessments, the meaning, interpretation, and analysis of linguistic features across different demographic groups is a pressing matter. This line of research is necessary to ensure fair assessments for all learners.

4.3 RQ3: Latent syntactic and lexical factors predicting reading outcomes

The third research question focuses on what latent factors can be identified in the NLP-derived lexical and syntactic features, using an unsupervised ML method, and then investigates how those factors predict reading comprehension using traditional multiple regression. This research question differs from RQ1 in that it concerns the relationship between reading comprehension and the latent factors underlying the lexical and syntactic features, rather than between reading comprehension and individual lexical and syntactic features as in RQ1. Methodologically, RQ1 and RQ3 are different, in that for RQ1, 260 individual lexical and syntactic features were modelled as individual variables in predicting reading comprehension (a one-step process), while for R3 a two-step process was used: first, an unsupervised ML approach identified the structure of latent factors underlying the NLP datasets, and then in a post-hoc fashion, used those factors to predict reading comprehension. This two-step approach, first using unsupervised ML to reduce the dimensionality of the data (also known as ML feature extraction), and then developing a predictive function that uses the reduced-dimension features (in this case, latent factors) to predict an outcome variable, is a common approach to investigating high-dimensional datasets (Lo, Rensi, Torng & Altman, 2018).

Factor analyses were run separately on each of the four datasets (oral/regular, oral/modified, text/regular, and text/modified). Scree plots were created using the 'psych' package in R in order to determine the appropriate number of factors. Subsequently, factor loadings and factor scores were generated using Scikit-learn's FactorAnalysis program. Factors are uncorrelated and were not rotated, as rotation is not available in Scikit-learn's program. Below, factor loadings > .5 are reported due to the large number of variables. The final step in this process is multiple linear regression of reading comprehension scores on the factor scores. Ultimately, the goal is to determine any of the factor scores, which were created independently of the reading comprehension scores (hence the term "unsupervised"), are predictive of reading comprehension.

Figures are provided below as an overview of the factor loadings for each dataset, demonstrating the direction of the loading for each linguistic feature type (grammatical constituent, grammatical complexity, vocabulary richness and range, vocabulary specificity, similarity, and

ambiguity, and vocabulary sentiment and affect). As mentioned above, due to the large number of variables, only loadings greater than .5, and less than -.5 are included in this discussion; however, when factor scores are used for the multiple regression in the next section, they include the full loadings.

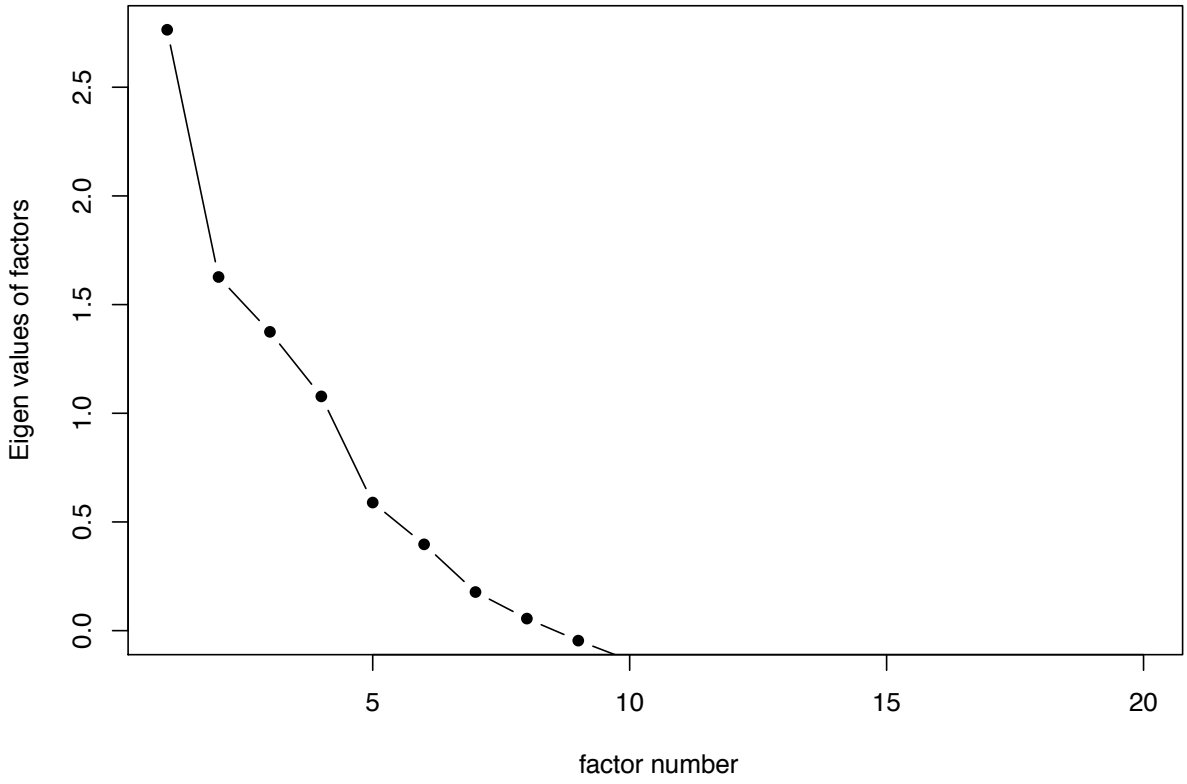

Figure 23. Scree plot for oral/regular dataset

Figure 23 shows the eigenvalues of each factor in the oral/regular dataset. Selection of the appropriate number of factors can be done visually by inspecting for an "elbow" in the plot, below which there is not significant variance absorbed by additional factors. Another method in common use is to use an eigenvalue cut-off, retaining all factors with eigenvalues greater than or equal to 1. In this case, four factors were retained, and Table 18 indicates the amount of variance explained by each factor. The four factors explain approximately one quarter of the total variance in the data. While this is not a very large proportion of variance explained, the dataset contains over 200 variables, so a model that explains approximately one quarter of the variance in such a large dataset with just 4 factors is certainly worth exploring further.

Table 18

Variance explained by exploratory factor analysis model of oral/regular dataset

	Variance explained by individual factor	Cumulative variance explained
Factor 1	.11	.11
Factor 2	.09	.20
Factor 3	.04	.24
Factor 4	.02	.26

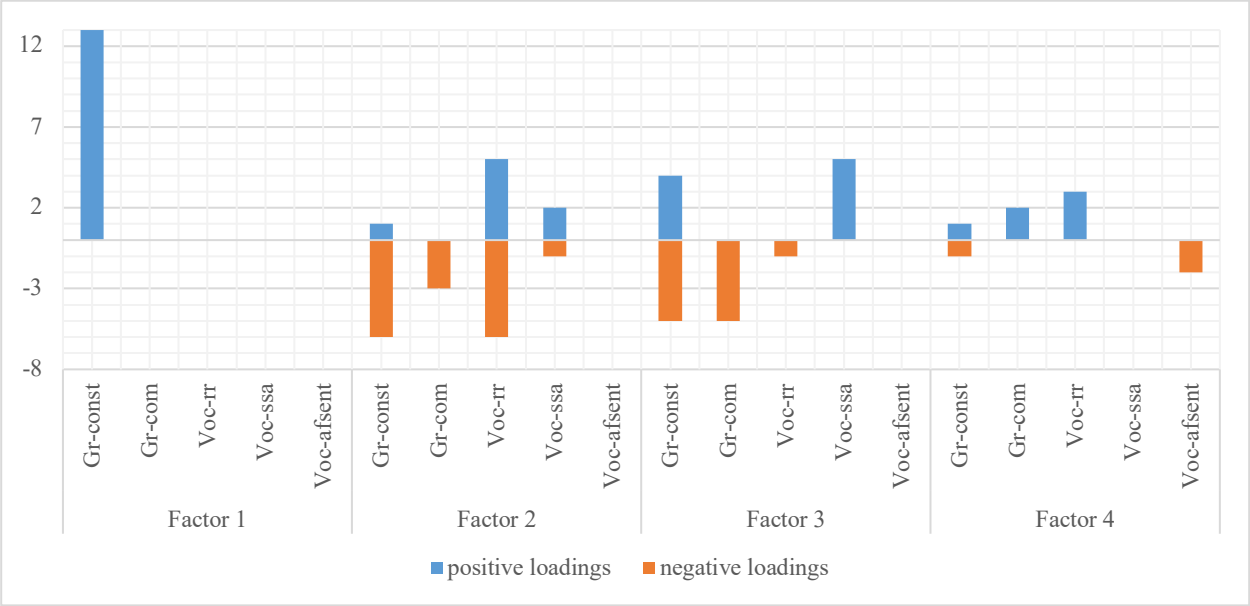

Figure 24. Number of variables loading at >|.5| for each factor in oral/regular dataset

Figure 24 illustrates the pattern of factor loadings. Factor 1 is solely grammatical constituents, Factor 2 is negative loadings of grammatical constituents and complexity, as well as both positive and negative loadings of vocabulary richness, with a few variables of vocabulary specificity, similarity, and ambiguity. Factor 3 is like Factor 2 except without positive loadings for vocabulary richness or negative loadings for vocabulary specificity, similarity, and ambiguity. Factor 5 is the only factor that includes affect and sentiment, which loads negatively, and all others but vocabulary specificity, similarity, and ambiguity. It is important to remember here that none of the variables have been reverse-coded, so some variables within categories may load onto factors in reverse. For example, low type-token ratio values are equivalent to high values in Brunet's Index, although both measure how many different words are used across the

length of a text. Table 19 further explores each factor by providing factor loadings > |.5| for these four factors.

Table 19

Factor loadings for oral/regular dataset, with negative loadings italicized

Factor 1	
Verb phrases per T-unit (Gr-com)	.99
Mean length of T-units (Gr-com)	.98
Complex T-units (Gr-com)	.95
Dependent clauses / T-unit (Gr-com)	.92
Complex nominal / T-unit (Gr-com)	.88
Complex phrases per T-unit (Gr-com)	.72
Length of verb phrases over length of sample (Gr-com)	.63
Average verb phrase length (Gr-com)	.63
Mean length of sentences (Gr-com)	.61
Clauses per sentence (Gr-com)	.60
Words (Gr-com)	.59
Average height of each parsed tree in the sample (Gr-com)	.56
Complex T-units per T-unit (Gr-com)	.53
Factor 2	
Brunet's Index (Voc-rr)	.89
Noun-verb ratio (Gr-const)	.80
Kurtosis of verb ambiguity (Voc-ssa)	.75
Age of acquisition (Voc-rr)	.71
Not in dictionary (Voc-rr)	.69
Noun age of acquisition (Voc-rr)	.66
Skewness of word ambiguity	.59
Average word length (Voc-rr)	.58
Particles (Gr-const)	*-.51*
Familiarity of nouns (Voc-rr)	*-.51*
Average word ambiguity	*-.52*
Roots (Gr-const)	*-.53*
Complex nominals (Gr-com)	*-.54*
Word familiarity (Voc-rr)	*-.58*
Prepositional phrase that includes a noun phrase (Gr-const)	*-.58*
Length of prepositional phrases over entire sample (Gr-com)	*-.59*
Noun phrase with determiner (Gr-const)	*-.61*
Word imageability (Voc-rr)	*-.63*
Simple declarative clauses (Gr-const)	*-.64*
Function words (Gr-const)	*-.66*
Word frequency (Voc-rr)	*-.71*
Length of noun phrases over entire sample (Gr-com)	*-.76*
Noun imageability (Voc-rr)	*-.77*
Type-token ratio (Voc-rr)	*-.84*
Factor 3	
Average maximum depth from word to hypernym root (Voc-ssa)	.61
Average minimum depth from word to hypernym root (Voc-ssa)	.59
Count of nouns (Gr-const)	.58
Variation in word similarity (WP method) (Voc-ssa)	.57
Average word similarity (Lin semcor method) (Voc-ssa)	.57
Verb phrase consisting of non-3rd person singular verb and noun phrase (Gr-const)	.53
Noun phrase consisting of noun phrase and verb phrase (Gr-const)	.53
Determiners (Gr-const)	.52
Variation in word similarity (Lin semcor method)	.51
Clause introduced by subordinating conjunction (Gr-const)	*-.50*
Mean length of sentences (Gr-com)	*-.51*
Words (Gr-com)	*-.51*
Propositional density (Gr-const)	*-.52*
Mean depth (Gr-com)	*-.53*
Moving average type-token ratio (window of 20 words) (Voc-rr)	*-.53*
Total depth (Gr-com)	*-.55*
Clauses per sentence (Gr-com)	*-.57*
Noun phrase with personal pronoun (Gr-const)	*-.70*
Count of Personal pronouns (Gr-const)	*-.75*

Ratio of personal pronouns to personal pronouns and nouns (Gr-const)	*-.77*
<u>*Factor 4*</u>	
Moving average type-token ratio (window of 50 words) (Voc-rr)	.62
Noun phrase → noun phrase and prepositional phrase (Gr-const)	.57
Moving average type-token ratio (window of 40 words) (Voc-rr)	.56
Moving average type-token ratio (window of 30 words) (Voc-rr)	.54
Complex nominal (Gr-com)	.54
Length of noun phrases over length of sample (Gr-com)	.53
Variation in noun arousal (Voc-afsent)	*-.50*
Mean arousal (Voc-afsent)	*-.50*
Verbs (Gr-const)	*-.53*

Confirming the plot in Figure 24, the first factor is entirely composed of grammatical complexity variables, and it is intuitive that they would be correlated and thus load onto a factor. The number of words, however, is not exactly complexity, so it is interesting to note that the length of the narratives in the oral/regular dataset is correlated with the grammatical complexity variables.

Factor 2 has positive loadings from vocabulary range variables (age of acquisition, word length), and not-in-dictionary words are also loading on this factor – because these data were transcribed by graduate students, these are likely not spelling errors but instead proper nouns. Vocabulary-related variables negatively loading on this factor align with those already discussed: familiarity, frequency, and imageability are lower when age of acquisition is higher. Interestingly, sophisticated vocabulary appears to be associated with less diverse word usage, as the negative loadings of type-token ratio (although its limitations have been discussed) and positive loadings of Brunet's index denote. Negative loadings of complex nominals and the length of noun and prepositional phrases could be interpreted as a trade-off for more sophisticated vocabulary. Average word ambiguity loads negatively, with familiarity and imageability, which does not align with the previous discussion for this factor. Skewness and kurtosis of word ambiguity load positively, suggesting the use of a few highly ambiguous words (rather than a high average of ambiguous words) aligns with age of acquisition. Function words and particles are both loading negatively for this factor as well, along with simple declarative clauses. In sum, Factor 2 primarily pertains to vocabulary sophistication, which appears to be negatively correlated with the length of noun and prepositional phrases (leading to the potential conclusion that longer phrases of these sorts are roundabouts when vocabulary cannot be conjured). Yet simple declarative clauses negatively correlating here also suggests that simple grammatical forms are negatively related to sophisticated vocabulary use.

Factor 3 is primarily positive loadings for vocabulary specificity and similarity, and negative loadings for grammatical complexity, with some grammatical constituents as well. The strongest loadings for Factor 3 are negative, specifically relating to personal pronoun usage. Several grammatical complexity items also loaded negatively on this factor, which may relate to personal pronoun use, since personal pronouns can act as referents for clauses. Total narrative length and moving type-token ratio (20-word window) also loaded in this direction. Propositional density, (which is the ratio of verbs, adjectives, adverbs, prepositions, and conjunctions to the total number of words) was also a negative loading. Positive loadings for Factor 3 included average word specificity and similarity (and similarity variation), noun constituents, and determiners. Overall, this factor is difficult to interpret, but it can be summarized as having positive loadings for noun usage and word similarity and specificity, and negative loadings for grammatical complexity, personal pronoun usage, and vocabulary diversity. If words are similar, then they are less likely to be diverse, which is intuitive. Interestingly, noun count is associated with shorter sentences, fewer clauses per sentence, and fewer overall words. If nouns are indeed learned before verbs, and are easier to learn than verbs, then this is also intuitive to interpret.

Factor 4 had positive loadings from several windows (30, 40, and 50 words) of the moving average type-token ratio, representing lexical diversity, as well as three variables relating to complex noun phrases. Negative loadings for Factor 4 included verbs and mean and variation in arousal (which is measured from calm to excited).

Table 20

Multiple regression model with factor scores predicting reading comprehension scores (oral/regular dataset)

Latent variable	*B*	*SE*	*T-value*	*P*
Factor 1	.36	1.20	.30	.76
Factor 2	-.24	1.21	-.20	.85
Factor 3	-3.73	1.22	-3.06	<.01**
Factor 4	-1.58	1.22	-1.30	.20
Intercept	74.45	1.20	62.04	<.01
R^2 (adjusted)		.07		
F		2.79 (4,90)		.03*

The regression results for the factor scores from the oral/regular data (Table 20) were significant and explained approximately 7% of the variance in the outcome variable (when rerun with only Factor 3, the total variance explained was 8.3%). The only individual factor achieving

significance is Factor 3, which negatively predicts reading comprehension. This means that this factor's positive loadings for noun usage and word specificity and similarity (and variation therein) are inversely related to reading comprehension, while the factor's negative loadings for grammatical complexity, total length, personal pronoun usage, and vocabulary diversity (across a short window) are positively associated with reading comprehension. Overall, this aligns with the literature on factors related to language and/or literacy development, except perhaps for word specificity, which one might expect would positively predict reading skill. However, as discussed above, the WordNet specificity metric is not necessarily a measure of lexical sophistication.

Since Factor 3 was a significant predictor of reading comprehension, it deserves a closer look, especially with regard to the WordNet features that were predominant in the positive loadings of this factor. Figure 25 shows the relationship between Factor 3 and reading comprehension for the oral/regular data. The plot on the left has all 95 observations present in the oral/regular dataset. There is a high leverage observation around positive 4 on the x-axis (Factor 3), tilting the fitted line to a steep negative slope. Another outlier is around negative 3 on the x-axis. These two observations represent the highest and lowest values of Factor 3, respectively.

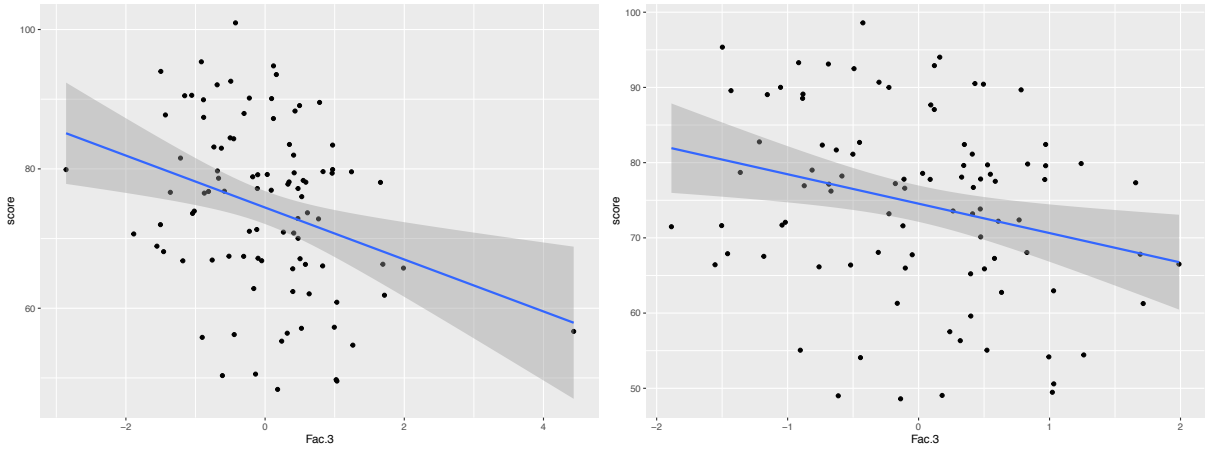

Figure 25. Relationship between Factor 3 and reading comprehension score in the oral/regular dataset. Left panel: with two high-leverage observations. Right panel: without two high-leverage observations

Examining the original narratives, it turns out that the +4 observation only has two narrative samples, and both are for picture description (this participant did not complete either story retell task). The observation at -3 only has one narrative, which is for SR2. In data preprocessing, features were averaged across the two oral task types (four tasks in total), but these two participants each lacked representation in one of the task types. It appears that the difference in task demands accounts for some of the differences in the features. For example, the story retell responses tend to be longer than the picture description responses, and number of words loaded negatively on Factor 3. (the word count for the +4 observation averaged 26.5 words, while the sole language sample provided by the -3 observation was 176 words long). The plot on the right of Figure 25 eliminates those two outliers, and the negative trend is still evident, although not as steep.

I examined the narratives that are not outlying on Factor 3 due to an artifact of task completion. Here are excerpts from Participant 915, who had the second-highest score on Factor 3, from tasks PD1 and SR2:

> In the picture there is a dog with their red ball two kids one is a girl one is a boy one mother grabbing an apple also blending fruits blending maybe a smoothie of some sort music there is music blaring from a stereo I see two cloths one is blue and white and one is red and pink there is a toaster with toast ready a girl is falling off a stool as she's trying to grab what looks like cupcakes there is a grey stove and a white counter top with purple and light blue pulls and [noise] (P915, PD1)

> Pablo and his friend [um] were practicing for the track and field team even though it was winter and a bully named Mac was taller and stronger than them even though they were in the same grade [unk] stopped in their path making it hard for them to race apparently Pablo and his friend or at least just Pablo gets bullied by Mac everyday and just this day when all his friend stood up for him telling the bully Mac that we were racing and he didn't want

> them in their way until they would leave him alone [unknown] friend that made Pablo happy and strong that his friend was on his side as Mac told him that he never made the track and field team cause he wasn't good at running so they all practiced together at the end of recess (P915, SR2)

Excerpts from Participant 838, who had the second-lowest scores on Factor 3, on the same tasks:

> I think it's like a family and they're in a kitchen and the mom is like making like a smoothie of some sort and then the little boy is playing with his dog while they're listening to music and the girl is because her mother put away the cupcake maybe she's trying to reach for them (P838, PD1)

> So there was these two friends Pablo and Ariana and they wanted to be on the school's track and field team so they were practicing this one recess and then there was a person blocking the way and then he saw it was Mac the school bully even though mac and Pablo were in the same grade Mac was like a lot taller and stronger than Pablo and then Mac said get out of my way loser and then Pablo wanted to say something but he couldn't because he was too scared and then and then his and then Ariana said like he's my friend and then Mac was like surprised that Ariana would stand up for Pablo and then she asked do you want to race with us … (P838, SR2)

With regard to differences in Factor 3, the first example (P915) uses several words that have a long path to the hypernym root in WordNet: "toaster" (depth of 11), "stereo" (minimum depth of 9, maximum depth of 10), "smoothie" (minimum depth of 7, maximum depth of 9), "path" (depth of 5). Proportionately, P915 also uses more determiners and has a higher count of nouns. These features were associated with high scores on Factor 3. The picture description narrative tends toward the "listing"-type of response that in previous analyses has been a negative predictor of reading comprehension. As for low scores on Factor 3, P383 uses very long

sentences, largely through the use of the word “and,” but also “so”, which accounts for the clauses introduced by subordinating conjunctions. Differences in pronoun use are not evidenced in these samples. In summary, Factor 3 is a significant negative predictor of reading comprehension in the oral/regular data. Examining the narrative samples with high and low factor scores allows for easier interpretation of the constructs represented in the factor. Caution must be exercised with regard to interpretation, as bias in a factor may be an artifact due to data pre-processing; in this case, two high leverage observations appears to have extreme scores for Factor 3 in part due to the two participants not completing any tasks of a certain type. Nonetheless, even with those two observations removed, the factor is still a negative predictor of reading comprehension, per visual inspection.

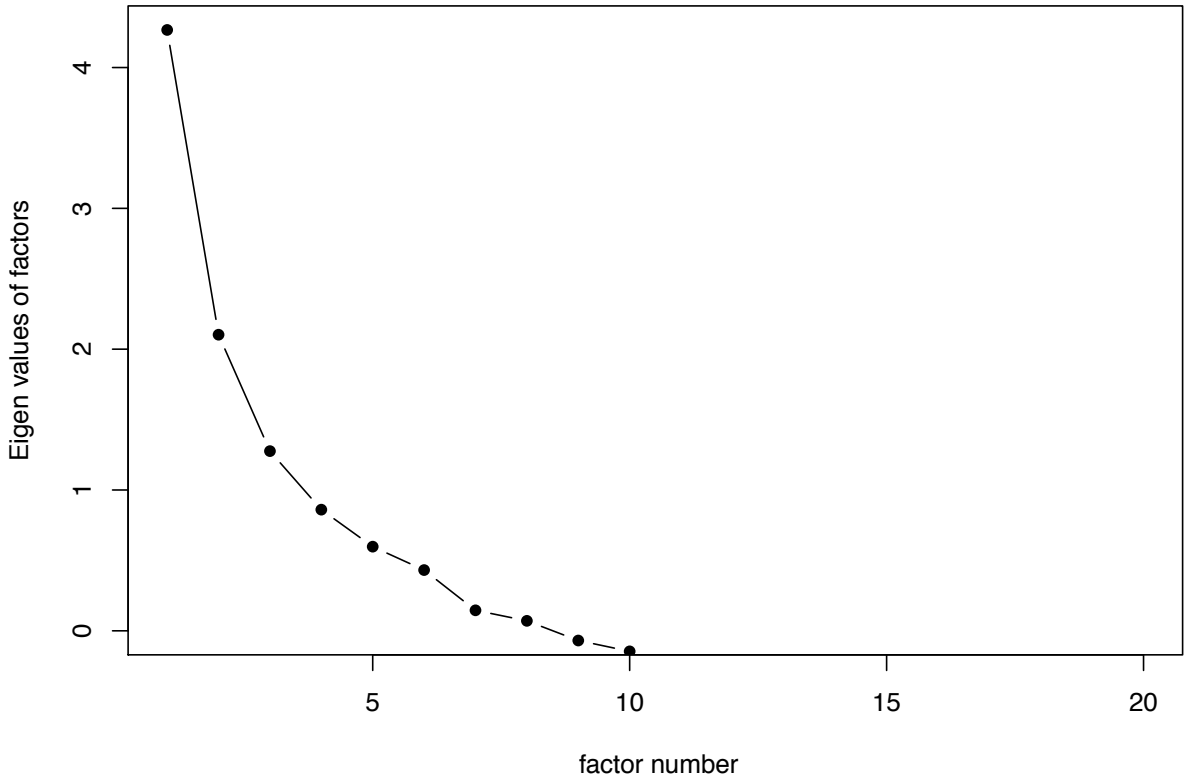

Figure 26. Scree plot for oral/modified dataset

Figure 26 shows the eigenvalues of each factor in the oral/modified dataset. In this case, three factors were retained, and Table 21 indicates the amount of variance explained by each factor.

Like the previous section, the retained factors explain approximately one quarter of the total variance in the data.

Table 21

Variance explained by exploratory factor analysis model of oral/modified dataset

	Variance explained by individual factor	Cumulative variance explained
Factor 1	.14	.14
Factor 2	.07	.21
Factor 3	.02	.23

Figure 27. Number of variables loading at >|.5| for each factor in oral/modified dataset

Figure 27 illustrates the pattern of factor loadings for the oral/modified dataset. Grammatical constituents have strong loadings for all three factors. In Factor 1, vocabulary specificity, similarity, and ambiguity have a strong negative loading, as well as vocabulary affect and sentiment. In Factor 2, vocabulary specificity, similarity, and ambiguity is again loading strongly negatively, this time with vocabulary richness and range; affect and sentiment load positively. Factor 3 is primarily composed of grammatical constituent variables. The factor loading patterns in Figure 27 do not share much in common with those in the previous dataset. For example, vocabulary affect/sentiment and vocabulary specificity, similarity, and ambiguity are more

prevalent in this model, and there is not a factor in which grammatical constituents are the only feature type. Table 22 provides a more in-depth look at the composition of these three factors.

Table 22

Factor loadings for oral/modified dataset, with negative loadings italicized

Factor 1	
Noun phrase consisting of determiner, noun, and noun (Gr-const)	.64
Particle (Gr-const)	.61
Roots (Gr-const)	.56
Prepositions (Gr-const)	.55
Variation in valence (Voc-afsent)	.53
Noun phrase consisting of personal pronoun and noun (Gr-const)	.52
Variation in verb valence (Voc-afsent)	.51
Valence mean noun (Voc-afsent)	*-.55*
Moving type-token ratio (window of 10 words)	*-.56*
Variation in word similarity (Resnick Brown method)	*-.57*
Mean dominance of nouns (Voc-afsent)	*-.59*
Mean dominance (Voc-afsent)	*-.62*
Mean valence (Voc-afsent)	*-.65*
Average word similarity (LC method) (Voc-ssa)	*-.65*
Average word similarity WP (Voc-ssa)	*-.77*
Average word similarity (Lin Brown method) (Voc-ssa)	*-.89*
Average word similarity Linsemcor (Voc-ssa)	*-.90*
Variation in word similarity (LC method)	*-.91*
Variation in word similarity WP (Voc-ssa)	*-.96*
Variation in word similarity (Lin Brown method) (Voc-ssa)	*-.99*
Variation in word similarity Linsemcor (Voc-ssa)	*-1.00*
Factor 2	
Roots (Gr-[c]onst)	.73
Verb phrase consisting of 3rd-person singular verb and verb phrase (Gr-const)	.68
Average maximum depth from noun to root hypernym (Voc-ssa)	.66
Type-token ratio (Voc-rr)	.64
Mean noun arousal (Voc-afsent)	.64
Noun phrase consisting of personal pronoun and noun (Gr-const)	.61
Imageability (Voc-rr)	.60
Variation in valence (Voc-afsent)	.59
Particle (Gr-const)	.59
Mean arousal (Voc-afsent)	.58
T-unit (Gr-com)	.57
Noun phrase consisting of a determiner, noun, and noun (Gr-const)	.57
Variation in noun valence (Voc-afsent)	.56
Variation in word dominance (Voc-afsent)	.55
Variation in verb dominance (Voc-afsent)	.52
Noun imageability (Voc-rr)	.50
Kurtosis ambiguity verb (Voc-ssa)	*-.53*
Variation in maximum depth from word to root hypernym verb (Voc-ssa)	*-.54*
Skewness ambiguity verb (Voc-ssa)	*-.55*
Variation in minimum depth from word to root hypernym verb (Voc-ssa)	*-.56*
Total sentence depth (with weighting for left-branching sentences) (Gr-com)	*-.59*
Mean dominance (Voc-afsent)	*-.60*
Noun age of acquisition (Voc-rr)	*-.60*
Variation in ambiguity noun (Voc-ssa)	*-.61*
Moving type-token ratio (window of 10 (Voc-rr)	*-.62*
Age of acquisition (Voc-rr)	*-.63*
Noun phrase consisting of a personal pronoun (Gr-const)	*-.67*
Clauses per sentence (Gr-com)	*-.67*
Moving type-token ratio (window of20 (Voc-rr)	*-.68*
Mean length of sentence (Gr-com)	*-.68*
Number of words (Gr-com)	*-.69*
Average ambiguity noun (Voc-ssa)	*-.69*
Mean verb dominance	***-.75***
Moving type-token ratio (window of 50 (Voc-rr)	*-.77*
Brunet's Index (Voc-rr)	*-.77*
Moving type-token ratio (window of 30 (Voc-rr)	*-.77*
Moving type-token ratio (window of 40 (Voc-rr)	*-.78*
Factor 3	

Noun ratio (Gr-const)	.86
Noun-verb ratio (Gr-const)	.76
Nouns (Gr-const)	.75
Not in dictionary (Voc-rr)	.72
Word length (Voc-rr)	.54
Length of verb phrases (Gr-com)	*-.50*
Clauses (Gr-com)	*-.51*

Verb phrase consisting of "to" and verb phrase (Gr-const)	*-.53*
Propositional density (Gr-const)	*-.58*
Verb phrase (Gr-const)	*-.65*
Function (Gr-const)	*-.66*
Frequency (Voc-rr)	*-.67*
Number of verb phrases (Gr-const)	*-.73*
Verbs (Gr-const)	*-.76*

The first factor in the oral/modified data has positive loadings from simple noun phrases, prepositions, participles, and roots, along with variation in valence (unpleasant to pleasant). The majority of the variables loading on this factor, however, are negative loadings pertaining to averages of, and variations in, word similarity, as well as mean dominance (controlled to in control) and valence. The variation in lexical diversity over a 10-word window is also negatively loading on this factor. This factor suggests that greater use of certain noun and preposition usage is associated with less word similarity (and less variation in that similarity), less lexical diversity, and more words that tend to be less pleasant and more "controlled". Put the opposite way, greater word similarity (and variation in that similarity) is associated with more pleasant and "in control" (dominance) vocabulary usage, greater lexical diversity across a short window of words, and less use of certain simple noun phrases, participles, and prepositions. This factor contradicts Factor 3 in the oral/regular model because here, word similarity and lexical diversity are positively correlated, although the moving window is only 10 words, whereas before it was 20 words in length.

Factor 2 in the oral/modified data has a large number of variables loading at > |.5|. The negative loadings were the strongest. These include all the moving type-token averages (with windows from 10-50 words), three variables measuring grammatical complexity, two age-of-acquisition variables, Brunet's Index, two mean dominance (controlled to in control) variables, mean (and all functions of) ambiguity, and variation in word specificity. Loading the other direction on this factor were the normalized count of roots and T-units (which usually inversely correlate to grammatical complexity), imageability (inversely correlating with age of acquisition), type-token ratio (inversely correlating with Brunet's Index), simple noun p^{hr}ases and particles, verb phrase with 3rd-person singular verb and verb phrase (which does not readily appear to relate to the

other loadings). Also loading positively here were mean word specificity and arousal, and variation in valence, arousal, and dominance. The affect-sentiment category loading onto this factor has mean arousal and variation in valence and dominance loading positively, while mean dominance is loading negatively. Like the previous factor, this one is difficult to interpret in some aspects. The vocabulary range features load as expected, as well as ambiguity. Vocabulary richness load in the same direction, as do complex grammatical features. The dominance and arousal features loading here are more difficult to interpret. Warriner et al. (2013) note in their discussion of the creation of the corpus, that there is a "positive correlation between dominance and arousal for high-rated dominance words…[and a] negative correlation between dominance and arousal for low-rated dominance words" (p. 1196). Thus, the loading of arousal against dominance is not entirely unexpected.

Factor 3 by comparison is rather straightforward to interpret. Positive loadings relate to nouns, not-in-dictionary words (likely proper nouns), and word length, while negative loadings relate to verbs, normalized count of clauses, function words, propositional density (the ratio of verbs, adjectives, adverbs, prepositions, and conjunctions to the total number of words) and frequency of vocabulary. It is somewhat surprising that the number of verb phrases and the length of verb phrases load onto the same factor, since those are usually inversely related. Additionally, given that verbs are harder to learn than nouns, it is somewhat expected that word length, a proxy for vocabulary range, would load against verbs, while vocabulary frequency (higher values meaning more frequent) would load with verbs. Given the nature of the oral tasks – especially picture description – nouns play an important role. Of interesting about this factor is the direct tradeoff between verb and noun use; being higher in one is associated with being lower in the other.

Table 23

Multiple regression model with factor scores (oral/modified model)

Latent variable	*B*	*SE*	*T-value*	*p*
Factor 1	1.56	1.50	1.04	.30
Factor 2	-2.46	1.51	-1.63	.11
Factor 3	-.29	1.53	-.19	.85
Intercept	81.43	1.50	54.33	<.01
R^2 (adjusted)		.01		
F		1.25 (6,63)		.30

The regression of the modified reading comprehension score on these three factors was not significant (Table 23). Factor 2 was the closest to significance (p=.11), and it was negatively related to reading comprehension. This aligns with existing research on the correlation between reading comprehension and complex grammar and sophisticated vocabulary usage, but the positive contribution of dominance, and variation in word specificity, was not expected.

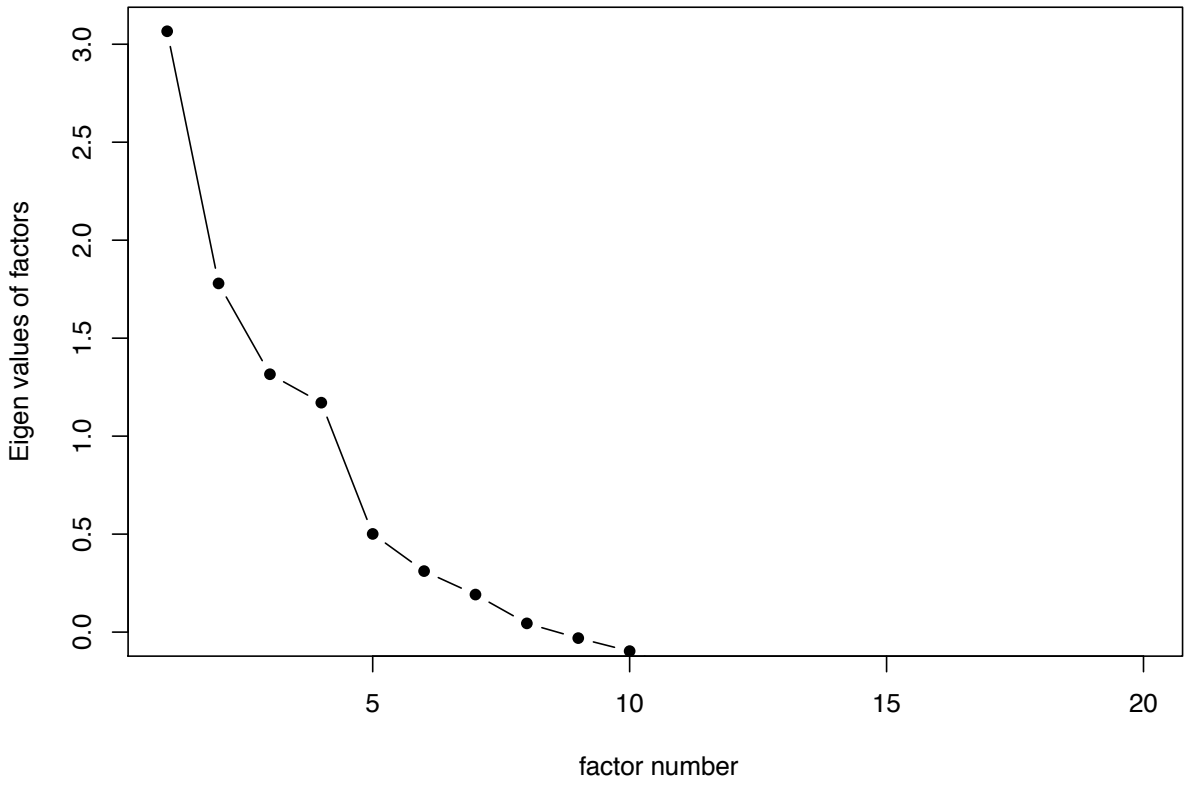

Figure 28. Scree plot for text/regular dataset

Figure 28 demonstrates the eigenvalues of each factor in the text/regular dataset, where four factors were retained. Table 24 indicates that less of the variance was explained by this factor analysis (14%) than in the previous datasets.

Table 24

Factor loadings for text/regular dataset

	Variance explained by individual factor	Cumulative variance explained
Factor 1	.07	.07

Factor 2	.03	.10
Factor 3	.02	.12
Factor 4	.02	.14

As Figure 29 demonstrates, there were fewer overall variables loading on these factors than in previous models. In Factor 1, grammatical complexity loads in the same direction as vocabulary richness and vocabulary specificity, similarity, and ambiguity. In Factor 2, it loads opposite vocabulary specificity, similarity, and ambiguity. Grammatical constituents form most of Factor 4. This result looks more similar to the oral/regular factor analysis result than the oral/modified factor analysis result. Interestingly, there were no loadings on any factor by the vocabulary sentiment and affect feature type.

Figure 29. Number of variables loading at >|.5| for each factor in text/regular dataset

Table 25 provides a close look at the factor loadings for the four factors.

Table 25

Factor loadings for text/regular dataset, with negative loadings italicized

Factor 1	
T-unit (Gr-com)	.80
Type-token ratio (Voc-rr)	.55
Mean sentence depth (with weighting for left-branching sentences) (Gr-com)	*-.53*
Variation in ambiguity noun (Voc-ssa)	*-.54*
Average word similarity (Lin Semcor method) (Voc-ssa)	*-.55*

Total sentence depth (with weighting for left-branching sentences) (Gr-com)	*-.55*
Max sentence depth (with weighting for left-branching sentences) (Gr-com)	*-.58*
Dependent clauses (Gr-com)	*-.60*
Moving type-token ratio (window of 50 (Voc-rr)	*-.60*
Clauses per sentence (Gr-com)	*-.60*
Prepositions (Gr-const)	*-.61*
Moving type-token ratio (window of 40 (Voc-rr)	*-.63*
Mean length of sentence (Gr-com)	*-.64*
Clauses/T-unit (Gr-com)	*-.65*
Verb phrases/T-unit (Gr-com)	*-.66*
Moving type-token ratio (window of 30	*-.67*
Complex nominals per T-unit (Gr-com)	*-.68*
Brunet's Index (Voc-rr)	*-.68*
Dependent clauses/T-unit (Gr-com)	*-.69*
Length of verb phrases (Gr-com)	*-.69*
Average height of each parse tree (Gr-com)	*-.69*
Complex T-units per T-unit (Gr-com)	*-.69*
Total number of words in each verb phrase divided by the total number of verb phrases	*-.70*
Variation in word similarity (Voc-ssa)	*-.72*
Mean length of T-unit (Gr-com)	*-.72*
Moving type-token ratio (window of 20 words (Voc-rr)	*-.74*
Dependent clauses per clause (Gr-com)	*-.75*
Variation in word similarity (LC method) (Voc-ssa)	*-.76*
Variation in word similarity (Lin Brown method) (Voc-ssa)	*-.76*
*Moving type-token ratio (window of 10 words (*Voc-rr)	*-.77*
Variation in word similarity (Lin Semcor method) (Voc-ssa)	*-.79*
<u>*Factor 2*</u>	
Clauses per T-unit (Gr-com)	.70
Verb phrases per T-unit (Gr-com)	.69
Average height of each parse tree (Gr-com)	.64
Dependent clauses/T-unit (Gr-com)	.64
Mean length of t-unit (Gr-com)	.63
Clauses per sentence (Gr-com)	.59
Complex nominal per T-unit (Gr-com)	.56
Length of noun phrases (Gr-com)	.55
Length of verb phrases (Gr-com)	.53
Mean length of sentence (Gr-com)	.51
Variation in word similarity (Lin Semcor method) (Voc-ssa)	*-.50*
Variation in word similarity (Lin Brown method) (Voc-ssa)	*-.53*
Variation in word similarity (LC method) (Voc-ssa)	*-.54*
Variation in word similarity wp (Voc-ssa)	*-.55*
Variation in word similarity (Resnick SemCor method) (Voc-ssa)	*-.56*
Variation in word similarity (Resnick Brown method) (Voc-ssa)	*-.59*
<u>*Factor 3*</u>	
Clauses (Gr-com)	.70
Verb phrases (Gr-const)	.61
Verbs (Gr-const)	.54
Number of words (Gr-com)	*-.51*
Moving type-token ratio (window of 40 (Voc-rr)	*-.52*
Mean length of clauses (Gr-com)	*-.60*
<u>*Factor 4*</u>	
Nouns (Gr-const)	.66
Noun-verb ratio (Gr-const)	.63
Average maximum depth from word to root hypernym (Voc-ssa)	.61
Average minimum depth from word to root hypernym (Voc-ssa)	.60
Noun ratio (Gr-const)	.57
Noun phrases consisting of a personal pronoun (Gr-const)	*-.50*
Ratio of Personal pronouns to Personal pronouns and nouns (Gr-const)	*-.59*
Frequency (Voc-rr)	*-.59*

Factor 1 is composed primarily of negatively loading factors against just two positively loading factors, which are the normalized count of T-units (usually inversely correlated with grammatical complexity) and type-token ratio. Indeed, the variables loading negatively here are measures of

grammatical complexity, moving type-token average (windows of 10, 30, 40, and 50 words), mean and variation of word similarity, variation in noun ambiguity, prepositions, and Brunet's Index. Factor 2 also has strong grammatical complexity loadings, this time in a positive direction, and all of the negative loadings are for variation in word similarity. Factor 3 has positive loadings for the count of clauses, verbs, and verb phrases, while negative loadings include the number of words, mean length of clauses, and moving type-token ratio (window of 40 words). Factor 4 has positive loadings from nouns and word specificity, and negative loadings from personal pronouns and word frequency. Interestingly, the only factor to have a vocabulary range feature is Factor 4, with vocabulary frequency. No factor for the text/regular data had a loading from vocabulary sentiment and affect.

Table 26

Multiple regression model with factor scores (text/regular model)

Latent variable	*B*	*SE*	*T-value*	*P*
Factor 1	-2.63	1.14	-2.34	.02*
Factor 2	-1.63	1.14	-1.43	.11
Factor 3	-3.04	1.15	-2.64	<.01**
Factor 4	2.49	1.16	2.15	.04*
Intercept	74.19	1.14	65.29	<.01
R2 (adjusted)		.13		
F		4.72	(4,94)	<.01

The regression of the regular reading comprehension score on these four factors was significant and explained approximately 13% of its variance (Table 26). Factors 1, 3, and 4 were all significant predictors (when rerun with only significant predictors, the total variance explained was 12.3%). Visuals of these are presented in Figure 30.

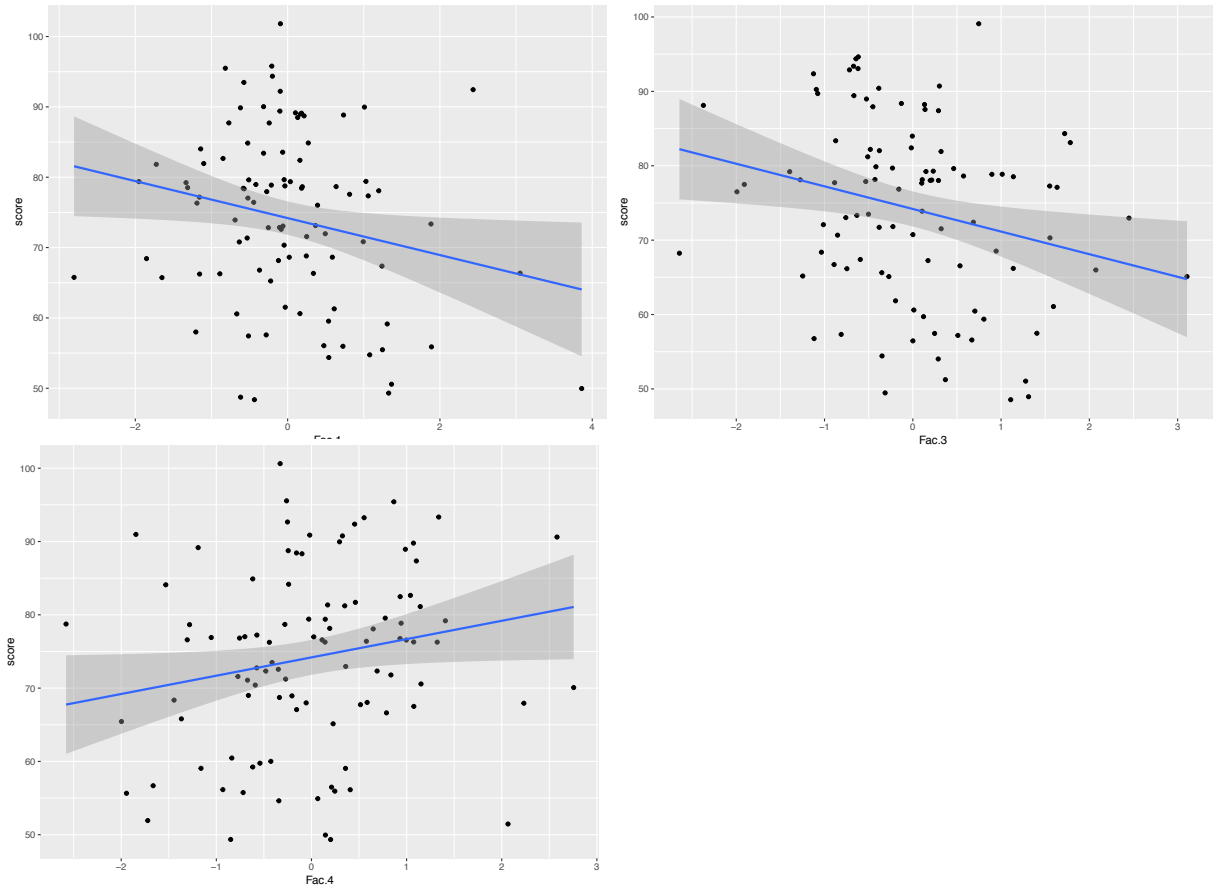

Figure 30. The relationship between factors produced through exploratory factor analysis and reading comprehension for text/regular data

Factor 1 was a significant negative predictor, which aligns with the knowledge base around complex syntax and lexical variation (in terms of moving type-token averages). However, it is somewhat surprising that Brunet's Index (greater lexical diversity for lower values) is negatively associated with reading comprehension, but it is an inverse to type-token average, which is loading positively. Again, the factor itself is a negative predictor, meaning the T-unit count and type-token-ratio that are positive loadings on Factor 1 *negatively* predict reading comprehension.

That word similarity positively predicts reading is a bit unexpected, but, again, it could relate to cohesion in writing. Participant 775 had the highest factor score for Factor 1, and this participant's sentences were exceptionally short, which explains the high count of T-units. In addition, each writing response was short (the entire writing response was "I don't know yet"),

accounting for the higher type-token ratio, which as discussed above is greater when total length is short. In addition, the negative loadings make sense for this participant's writing, such as dependent clauses per clause (virtually none), mean length of T-unit (very short), and what would be zero scores for moving average type-token ratios, since there are not sufficient words in the responses. On the other hand, a low score on Factor " in this dataset has complex clauses:

> "I think that games like candy rush for example are very unhealthy for your brain because whenever you complete a level or get an achievement your brain releases a chemical called dopamine which is what makes you happy, so candy crush a game that is designed to be adictive that also makes you happy sounds not so bad but heres the bad part when you play candy crush and it makes you feel good about yourself it messes up your brain so that only candy crush will make you happy. Now you can only be happy when you play candy crush, that's the same thing that drugs do to your brain but without the negative physical effects."

In addition, the variation in word similarity is evident as the writing is peppered with words like "chemical," "dopamine," and "physical", while staying focused on the topic at hand, which may account for the variation in word similarity.

Factor 3 had a significant negative relationship with reading comprehension, suggesting that clause length and total words are positive predictors, while use of verbs and total number of clauses operate in reverse. Lastly, Factor 4's nouns and word specificity are positively predicting reading comprehension, while the use of personal pronouns and higher frequency words are negative predictors. Overall, these findings align with the literature base, except the finding the verbs are negatively associated with higher reading outcomes (but again, this may be a task-specific result).

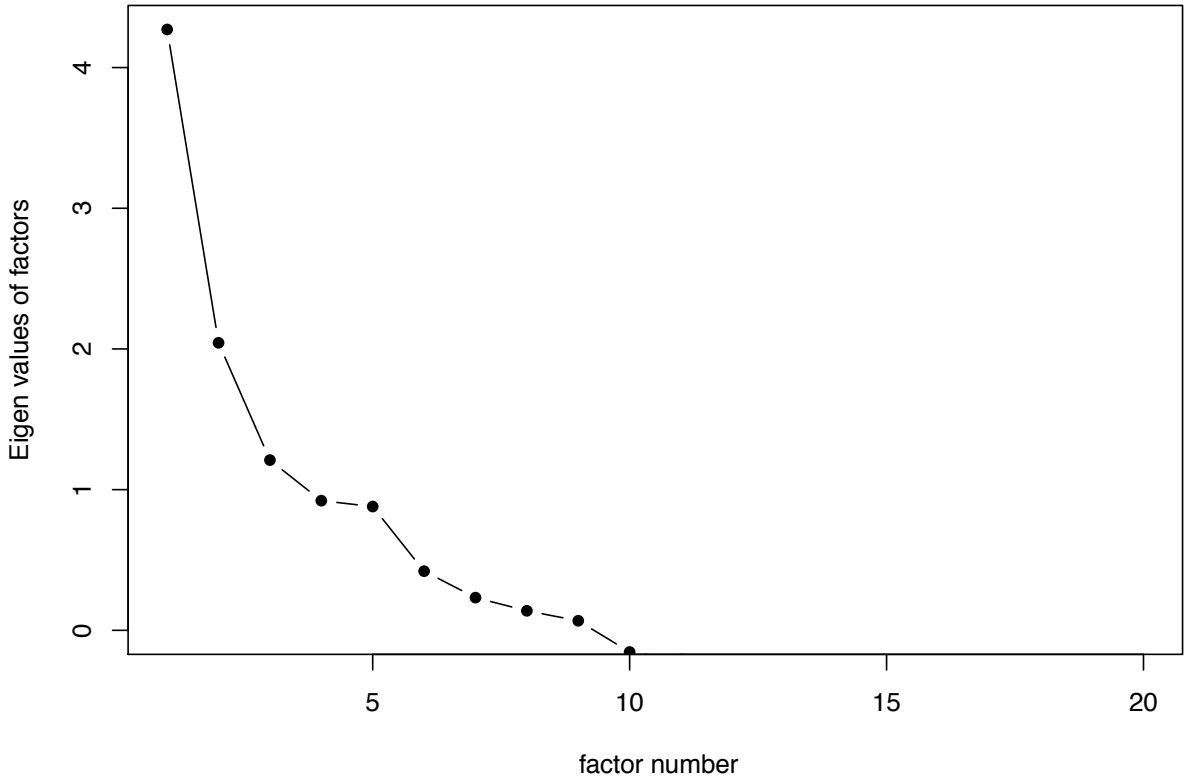

Figure 31. Scree plot for text/modified dataset

Figure 31 demonstrates the eigenvalues of each factor in the text/modified dataset, where five factors were retained. Table 27 indicates that 28% of the variance in the data was explained by this factor analysis.

Table 27

Factor loadings for text/modified dataset

	Variance explained by individual factor	Cumulative variance explained
Factor 1	.04	.04
Factor 2	.02	.06
Factor 3	.07	.13
Factor 4	.02	.15
Factor 5	.13	.28

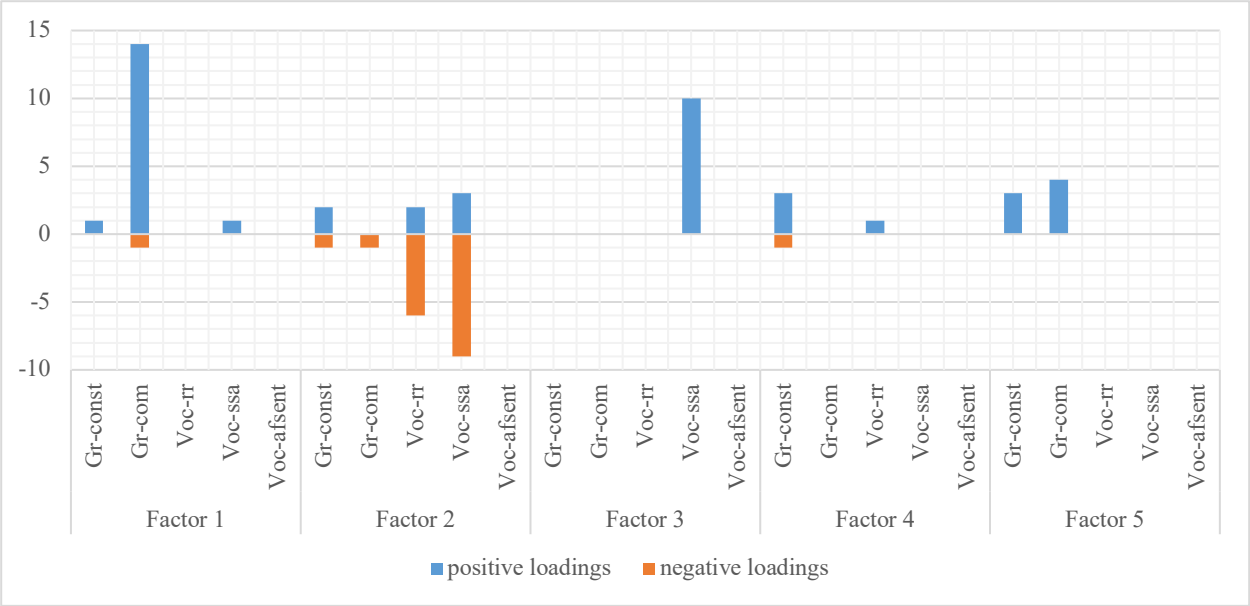

Figure 32. Number of variables loading at >|.5| for each factor in text/modified dataset

Similar to the factor analysis of the text/regular dataset, fewer overall variables in the text/modified data loaded onto each of these factors than for the oral-elicited data (Figure 32). Table 28 provides details about the factor loadings.

Table 28

Factor loadings for text/modified dataset, with negative loadings italicized

Factor 1	
Mean length of t-unit (Gr-com)	.99
Verb phrases per T-unit (Gr-com)	.98
Clauses per T-unit (Gr-com)	.98
Clauses per sentence (Gr-com)	.97
Mean length of sentence (Gr-com)	.96
Total sentence depth (with weighting for left-branching sentences) (Gr-com)	.94
Complex nominal per T-unit (Gr-com)	.92
Max sentence depth (with weighting for left-branching sentences) (Gr-com)	.91
Dependent clauses/T-unit (Gr-com)	.87
Average height of each parse tree	.87
Mean sentence depth (with weighting for left-branching sentences) (Gr-com)	.81
Length of verb phrases (Gr-com)	.72
Total length of verb phrases divided by the number of verb phrases (Gr-com)	.70
Prepositional phrase (Gr-const)	.55
Complex T-units per T-unit (Gr-com)	.55
Kurtosis of noun ambiguity (Voc-ssa)	.50
T-unit (Gr-com)	*-.70*
Factor 2	
Type-token ratio (Voc-rr)	.80
Root consisting of direct question introduced by a wh-word or a wh-phrase (Gr-const)	.72
Mean Stanford sentiment neutral	.68
Average maximum depth from word to root hypernym noun (Voc-ssa)	.66
Average minimum depth from word to root hypernym noun (Voc-ssa)	.57

Verbs (Gr-const)	.51
Noun imageability (Voc-rr)	.51
Variation in minimum depth from word to root hypernym noun (Voc-ssa)	*-.54*
Variation in ambiguity noun (Voc-ssa)	*-.55*
Variation in ambiguity verb (Voc-ssa)	*-.57*
Skewness of verb ambiguity (Voc-ssa)	*-.66*
Kurtosis of verb ambiguity (Voc-ssa)	*-.66*
Simple declarative clauses (Gr-const)	*-.67*
Variation in maximum depth from word to root hypernym verb (Voc-ssa)	*-.71*
Variation in minimum depth from word to root hypernym verb (Voc-ssa)	*-.71*
Kurtosis of word ambiguity (Voc-ssa)	*-.73*
Number of words (Gr-com)	*-.73*
Skewness of word ambiguity (Voc-ssa)	*-.78*
Moving type-token ratio (window of 10 (Voc-rr)	*-.81*
Moving type-token ratio (window of 50 (Voc-rr)	*-.82*
Brunet's Index (Voc-rr)	*-.88*
Moving type-token ratio (window of 40 (Voc-rr)	*-.89*
Moving type-token ratio (window of 30 (Voc-rr)	*-.90*
Moving type-token ratio (window of 20 (Voc-rr)	*-.90*
<u>*Factor 3*</u>	
Variation in word similarity (Lin Brown method) (Voc-ssa)	.95
Variation in word similarity (LC method) (Voc-ssa)	.91
Variation in word similarity (Lin Semcor method) (Voc-ssa)	.91
Variation in word similarity wp	.90
Average word similarity (Lin Semcor method) (Voc-ssa)	.76
Average word similarity wp (Voc-ssa)	.74
Average word similarity (Lin Brown method) (Voc-ssa)	.74
Variation in word similarity (Resnick Brown method) (Voc-ssa)	.69
Variation in word similarity (Resnick SemCor method) (Voc-ssa)	.68
Average word similarity (LC method) (Voc-ssa)	.63
<u>*Factor 4*</u>	
Noun ratio (Gr-const)	.93
Noun-verb ratio (Gr-const)	.88
Nouns (Gr-const)	.79
Not-in-dictionary (Voc-rr)	.63
Verbs (Gr-const)	*-.55*
<u>*Factor 5*</u>	
Coordinate phrases per T-unit (Gr-com)	.74
Coordinate phrases per clause (Gr-com)	.71
Coordinate phrases (Gr-com)	.64
Coordinates (Gr-const)	.61
Mean length of clauses (Gr-com)	.61
Verb phrase consisting of Verb phrase, coordinating conjunction, and verb phrase (Gr-const)	.54
Noun phrase consisting of noun phrase and verb phrase (Gr-const)	.53

Factor 1 has strong loadings by grammatical complexity variables. Negatively loading here is T-units, which is expected in light of all the grammatical complexity variables loading in the other direction. The strongest loadings in Factor 2 are negative. These are primarily functions (variation, skewness, and kurtosis) of word specificity and ambiguity and all the moving type-token average windows, as well as Brunet's Index. Again, the direction of Brunet's loading is surprising since lower values are for higher lexical sophistication. The number of simple declarative clauses and number of words were also negative loadings. As for positive loadings,

these were type-token ratio, average word specificity, the use of verbs, roots consisting of a question beginning with a *wh-* word, noun imageability, and neutral sentiment. Verbs, which previous analysis had shown to be positively associated with reading success, are here correlated with noun imageability. Factor 2 does not invite intuitive interpretation.

Factor 3 had all positive loadings of mean and variation in word similarity. Factor 4 consisted of nouns, not-in-dictionary words, and negative loading of verbs, again showing the verb-noun trade-off. Factor 5 consisted entirely of positive loadings, mostly relating to the use of coordinates (e.g., "and", "but"), but also the mean length of clauses and simple noun phrases that include a noun and verb phrase. This factor relates to simple, but long, clauses, which serves as a reminder that length of clause is not necessarily equivalent to clause complexity.

Table 29

Multiple regression model with factor scores (text/modified model)

Latent variable	*B*	*SE*	*T-value*	*P*
Factor 1	1.30	1.48	.88	.38
Factor 2	2.55	1.49	1.71	.09
Factor 3	-4.19	1.49	-2.81	<.01**
Factor 4	-.66	1.50	-.44	.66
Factor 5	-.31	1.51	-.21	.84
Intercept	81.70	1.48	55.07	<.01
R2 (adjusted)		.09		
F		2.40 (5,61)	<0.05	

The regression of the modified reading comprehension scores on these five factors was significant, as seen in Table 29 and explains approximately 9% of its variance. The individual factor with a significant relationship to the outcome variable was Factor 3, which is a negative predictor (variance explained solely by this factor, when rerun as simple linear regression, was 5.2%). This finding is not unexpected, as overuse of similar words can create repetitive writing and does not require as much cognitive effort as employing a more diverse lexis. However, as seen earlier, it is possible to effectively use high similarity among words to build cohesion. Possibly the participants who contributed to the modified dataset are still developing the control necessary to navigate that difference.

Figure 33 demonstrates the relationship between Factor 3 and reading comprehension in the text/modified dataset. Like before, there is a high-leverage observation above positive 4 on the factor score (plot on left panel of Figure 33). This participant (P745) did not complete the writing

task, and the results appear to be biased. The plot on the left panel of Figure 33 has that observation removed, and the predictive trend remains negative, although not as strongly so.

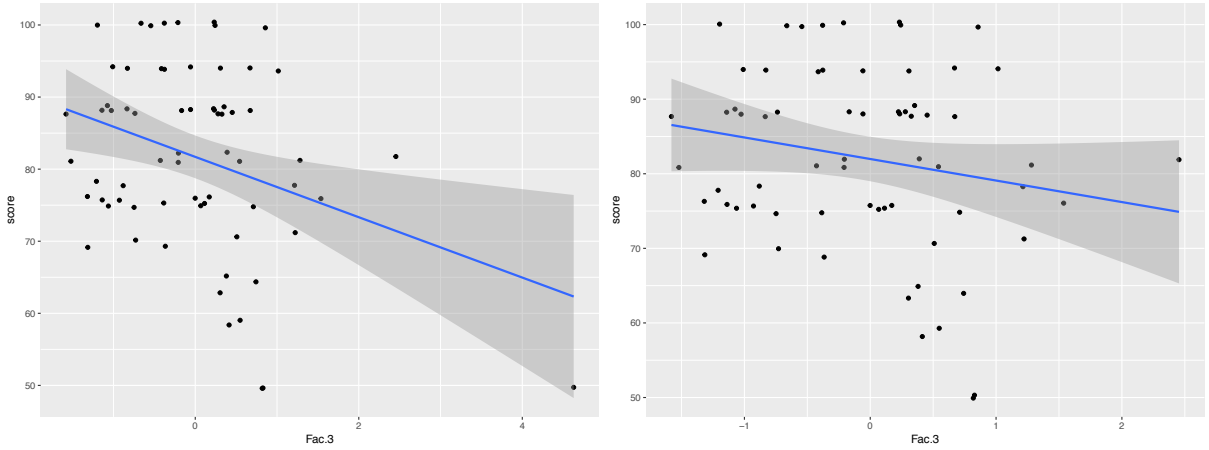

Figure 33. Relationship between Factor 3 (x-axis) and reading comprehension score (y-axis) for text/modified data; with a high leverage observation included (left) and excluded (right)

Two writing samples illustrate the difference between high and low scores on Factor 3. The first, P793, has a high factor score for Factor 3. The use of similar words is evident, with the repetition of "social media", "gadgets," and "you"; in addition, there are words that are similar but not identical such as "die", "sleep,", and "addicted" (these have similarity metrics between .22-.25). Yet there is also high variation between words, like "die" and "gadget" which are less semantically related:

> Social media is bad for you because you get addicted and you play games instead of being with your family. Social media is bad for your eyesight because so many people got glasses just for using gadgets. So many people died Just by gadgets, when they keep using gadgets while it's charged. Students got affected by social media because they sleep late because of social media and they forget to do their home works.

Participant 894 has a low factor score for Factor 3. There is some similarity amongst words in this sample, but less than the previous example since a variety of concepts are included:

> "I storngly belive that young people should not use social media because, it is horrible for eye sight, cyber bullying and less interaction. First of all, if you have too much time on screen, your eyes sight can go real bad. Secondly, cyber bullying. Let's say that George is spending time on his phone, than he recives a message from a stranger full of bad, swear words. Also, in these days now people can hack. If someone hacks you, they can take your personal information and delete everything! Last but not least, less interaction. Kids happen to be stuck on their phones and to not interact in social media. (eg. snap chap, insagram). etc)"

These writing samples illustrate how extensive use of similar words may be associated with lower writing quality and therefore an intuitive connection can be made between this and lower reading comprehension scores.

4.3.1.1 RQ3 Discussion

The research question guiding RQ3 focused on the use of ML-based factor analysis to identify the underlying factors of the four high-dimensional, NLP-derived lexical and syntactic feature sets. Unsupervised ML does not use labels to train a model; rather, it seeks patterns that are inherent in the independent data matrix. Factor analysis with Scikit-learn was the modeling algorithm of choice due to the lack of assumption of equal errors across variables. In turn, the resulting factors were entered into multiple regression models to determine if there was any predictive relationship between them and reading comprehension.

In these four datasets, between 3 and 5 factors were modeled depending on inspection of the scree plots, and the resulting factors explained between 14 and 28 percent of the variance within each dataset. A cut-off of an absolute value of .5 was set for factor loadings to report. Some factors were straightforward to interpret, but some were unwieldy with many seemingly unrelated features loading together. Nonetheless, none of the factors were so contradictory as to not make sense; rather, they combine some features that the existing literature does not address, and this makes interpretation a challenge.

Overall, the pattern of feature loadings showed that grammatical complexity was a stronger organizing construct for the BALA-regular datasets than the BALA-modified datasets. The vocabulary specificity, similarity, and ambiguity features did load onto several factors, but their interpretation remains a challenge. Of note was the relative absence of sentiment and affect features among the factor analyses, except the oral/modified data and very slightly in the oral/regular data.

With regard to the regression analyses, in the oral/regular dataset a factor composed primarily of grammatical complexity and vocabulary richness significantly and positively predicted reading comprehension. This factor also appeared to represent some "listing"-type language that negatively predicted the reading comprehension scores. Three of four factors in the text-regular data were predictive of reading comprehension: one, which was comprised of grammatical complexity and vocabulary richness, a second composed of clause length and total length, and a third which included vocabulary range and word specificity. These largely align with the existing literature base. In the text/modified data, a factor comprised solely of word similarity (and variation therein) was a significant negative predictor of reading comprehension, suggesting that repetition in writing can be negatively associated with reading outcomes. However, word similarity is not strictly associated with lower outcomes; further research is needed to understand the relationship. The factors underlying the oral/modified dataset did not significantly predict reading comprehension.

Tables 30-32 provide a different view of the factor analyses by organizing the features that loaded with, and against, grammatical complexity (Table 30), vocabulary range (Table 31), and vocabulary richness (Table 32) across the four different factor analysis models. Not all factors are recorded in these tables. The criteria for inclusion in these tables are that the factor has loadings from more than one language feature type (e.g., grammatical complexity) and that the factor has loadings from types that have been explored in existing research: grammatical complexity, vocabulary range, and vocabulary richness. In the text/modified dataset, for example, factors 3-5 are not included in these tables because they do not contain loadings of any of these three types. Factor 3 is just word similarity, Factor 4 is nouns and verbs, and Factor 5 is about length, but not complexity, of clauses. The goal of these tables is to better understand the functioning of features such as vocabulary specificity, similarity, and ambiguity, and vocabulary

sentiment and affect in relation to better-understood constructs (i.e., grammatical complexity, vocabulary range).

Table 30

Summary of features loading with (positively) and against (negatively) grammatical complexity features across the four models

Oral/ regular	Oral/ modified	Text/ regular	Text/ modified
Positive correlates			
Total length (F1) Total length Propositional density Personal pronouns Voc. richness (F3)	*[Factor 2 had loadings by sentence/clause length features but not sentence/clause Complexity features]*	Prepositions Voc. richness Brunet's Index Voc. ambiguity (fn) Voc. similarity (mn, fn) (F1)	Prepositions Voc. ambiguity (fn) (F1)
Negative correlates			
Nouns/noun phrases Determiners Voc. specificity Voc. similarity (mn, fn) (F3)		T-units Type-token ratio (F1) Voc. similarity (fn) (F2)	T-units (F1)

Note: F: Factor; mn: mean; fn: function; grammatical constituents are normalized counts; underlined features are replicated across datasets

Table 30 focuses on features that loaded positively and negatively with grammatical complexity features. Oral/modified is not included in this table because although Factor 2 in that dataset did have loadings related to grammar, these loadings were not specifically complex; they were just long. Because the dataset was elicited orally, there was a tendency for some participants to use "and" repeatedly. Recall Figure 18 which illustrates the oral/modified dataset having the greatest use of coordinating conjunctions; thus, it makes sense that this emerged as a factor. However, only having features relating to sentence/clause length – rather than complexity – does not justify that factor being included in this grammatical complexity table, which focuses on dependent clauses, clauses per T-unit, and so forth.

Table 30 provides several important pieces of information. First, although the factors that emerged from the four factor analyses are difficult to interpret, this table suggests that there are

some correlations that transcend datasets. Without the contribution of oral/modified, it is difficult to generalize about the modified datasets, but some comparisons can be made. Vocabulary richness (e.g., moving average type-token ratio) loads in the same direction as grammatical complexity for both regular datasets. This suggests that grammatical complexity and vocabulary richness are related in the regular datasets, but that in the modified datasets they do not correlate strongly.

In both text-based datasets, prepositions and functions (e.g., standard deviation) of vocabulary ambiguity load the same direction as grammatical complexity, and number of T-units (usually the inverse of complex grammar) load the opposite direction. The ambiguity functions can be interpreted as a high variation in the number of senses of the words used. In examining the ambiguity functions in the text datasets, it is not readily apparent how the variation in ambiguity reflects language or cognitive ability; this certainly merits further research. Functions of vocabulary similarity appear to load in two directions: both with and against grammatical complexity in the text/regular dataset; it also loads against it in the oral/regular set. Notably, although vocabulary richness loads with grammatical complexity, vocabulary range (e.g., age of acquisition) is not loading with or against grammatical complexity, nor are sentiment or affect features.

Table 31

Summary of features loading with (positively) and against (negatively) vocabulary range (positive: age of acquisition, word length; negative: frequency, imageability, familiarity) features across the four models

Oral/ regular	Oral/ modified	Text/ regular	Text/ modified
Positive correlates			
Nouns Brunet's index Voc. ambiguity (fn) (F2)	Sentence/clause length Voc. richness Voc. dominance Voc. ambiguity (mn, fn) Voc. specificity (fn)(F2) Nouns (F3)	Nouns Voc. specificity (F4)	Simple decl. clauses Total length Brunet's Index Voc. specificity (fn) Voc. ambiguity (fn) (F2)
Negative correlates			

Roots	Roots	Personal pronouns	Verbs
Simple decl. clauses	T-units	(F4)	Type-token ratio
Noun phrases with	Particles		Vocab. richness
determiners	Noun phrases		Questions beginning
Function words	Type-token ratio		with "wh-"
Particles	Voc. specificity		Neutral sentiment
Complex nominals	Voc. arousal		Voc. specificity
Length of noun and	Voc. valence (fn)		
prepositional phrases	Voc. dominance (fn)		
Type-token ratio	(F2)		
Voc. ambiguity (F2)			
	Verbs		
	Function words		
	Propositional density		
	Clauses (F3)		

Note: F: Factor; mn: mean; fn: function; grammatical constituents are normalized counts; underlined features are replicated across datasets

Table 31 summarizes features loading with and against vocabulary range features, which here include age of acquisition and word length, or negatively as imageability, familiarity, or frequency. In all but the text/modified data, noun use loaded with vocabulary range. This make intuitive sense given that nouns have a greater range of sophistication, in general, than verbs. Interestingly, mean vocabulary specificity loads against the vocabulary range for both modified datasets, suggesting that range and specificity are not positively correlated, as discussed above. However, it does load in the same direction for text/regular, suggesting that perhaps the lexical corpora trained on adults may be more accurate for the regular datasets than the modified datasets. Adding to the complexity is that the function of vocabulary specificity loads positively with vocabulary range for both modified datasets.

Factors with vocabulary range loadings did not have loadings relating to grammatical complexity, nor vocabulary richness, except for oral/modified. That "roots" were loading against vocabulary range for both the oral datasets suggests there may be fewer total sentences when higher vocabulary range is used; however, the complexity of those sentences is not loading here. Function words are also loading negatively, which makes sense considering their lexical level. Type-token ratio is loading against vocabulary range for all but text/regular, which provides further evidence that the interpretation of the standard type-token ratio remains a challenge.

Table 32

Summary of features loading with (positively) and against (negatively) vocabulary richness (moving average type-token ratio) features across the four models

Oral/ regular	Oral/ modified	Text/ regular	Text/ Modified
Positive correlates			
Gram. complexity Total length Propositional density Personal pronouns (F3, *loaded with MATTR20*) Complex noun phrases (F4, *loaded with MAtTR30-50*)	Voc. similarity (mn, fn) Voc. valence Voc. dominance (F1, *loaded with MATTR10*) Sentence/clause length Voc. Range Brunet's Index Total length Voc. range Voc. Dominance Voc. ambiguity (mn, fn) Voc. specificity (fn) (F2, *loaded with MATTR10-50*)	Prepositions Gram. complexity Voc. ambiguity (fn) Voc. similarity (fn) (F1, *loaded with MATTR10-50*) Total length Clause length (F3, *loaded with MATTR40*)	Simple decl. clauses Total length Brunet's Index Voc. specificity (fn) Voc. ambiguity (fn) (F2, *loaded with MATTR10-50*)
Negative correlates			
Nouns, noun phrases Determiners Voc. specificity Voc. similarity (mn, fn) (F3, *loaded with MATTR20*) Verbs Voc. arousal (mn, fn) (F4, *loaded with MATTR30-50*)	Particles Roots Nouns Prepositions Voc. valence (fn) (F1, loaded with MATTR10) Roots T-units Particles Noun phrases Voc. specificity Voc. valence (fn) Voc. dominance (fn) Voc. arousal (F2, *loaded with MATTR10-50*)	T-units Type-token ratio (F1, *loaded with MATTR10-50*) Clauses Verbs (F3, *loaded with MaTTR40*)	Verbs Type-token ratio Vocab. range Questions beginning with "wh-" Neutral sentiment Voc. specificity (F2, *loaded with MATTR10-50*)

Note: F: Factor; mn: mean; fn: function; grammatical constituents are normalized counts; underlined features are replicated across datasets

Table 32 summarizes features that loaded (either positively or negatively) with vocabulary richness (lexical diversity). As noted earlier, for both regular datasets, grammatical complexity and vocabulary richness are positively correlated. Vocabulary richness is also correlated with

total length in all four datasets, as well as with functions of vocabulary specificity and ambiguity for all but oral/regular. Verbs load against vocabulary richness in all but oral/modified; however, nouns are also negatively correlated with vocabulary richness in both oral models. There appeared to be some differences between features loading with the smaller and larger windows of the moving average type-token ratio. For example, nouns tended to load against moving type-token average with smaller windows (10-20 words), while verbs tended to load against windows of a larger size (30-50 words). This means that high diversity in a small window uses fewer nouns, while high diversity in a large window uses fewer verbs. Following is an example of the response to Story Retell 1 by a participant (P790) who completed the modified BALA assessment. This narrative has one of the lowest moving type-token average (window of 10 words) scores, and indeed, there is repetition within 10-word windows and an abundance of nouns, which are underlined:

> she asked her dad could you please drop me to school and then her dad's car was she felt embarrassed her dad's car was out of gas so when he was going to school he saw a gas station and then he he he filled it up and then after they reach school after they reached school Suzanne felt embarrassed because she was seven minutes late on her first day of school

The exploratory factor analyses described here support a deeper understanding of how the NLP-derived lexical and syntactic features map onto factors, and if those factors predict reading comprehension. It was found that the factors that emerge are interpretable, with some effort. Canonical lexical and syntactic constructs that educational researchers have investigated for decades, namely grammatical complexity and vocabulary richness and range, remain the easiest to interpret. It was found that the constructs of vocabulary richness (lexical diversity) correlate positively with grammatical complexity for the regular dataset, but vocabulary range (i.e., breadth and depth) does not appear to correlate strongly with either of these — except in the oral/modified model, where vocabulary richness and range loaded together.

In all but the oral/modified dataset, one or more of the factors underlying the lexical and syntactic feature dataset did significantly predict reading comprehension. These significant

predictors were largely composed of grammatical complexity and vocabulary diversity features (with one positively predicting factor composed of vocabulary range features, one of specificity, and one negative predicting factor composed of word similarity metrics). Again, use of high-similarity words appears to have mixed associations with successful reading comprehension: the association is positive for the regular data and negative for the modified data. Overall, the results across domains are similar, which supports the original hypothesis that the lexical and syntactic features of productive language, whether in written or spoken form, share similarities in predicting reading comprehension.

The factor analyses appeared to be exceptionally sensitive to task type, with outliers produced for participants who were missing responses to entire tasks (e.g., did not complete either picture description task). High leverage observations must to be interrogated to determine if their outlier status is due to a measurement or research design issue or if they are simply outside the normal distribution. Using data that is highly structured, that is, where responses are of similar lengths and data are guaranteed to be complete, can help ensure the validity of NLP-focused analyses. Nonetheless, the results here illustrate that NLP can provide unique insights into the relationship among language features in different domains and for different student populations.

5 General Discussion

The present study aimed to understand whether productive syntactic and lexical linguistic features, extracted through an out-of-the-box (meaning no modifications were made) NLP toolkit, predict reading comprehension for linguistically diverse children in grades 4-6. The goal of the research was to contribute to theoretical discussions surrounding the relationship between oral language and reading comprehension through a novel analytical approach that incorporates natural language processing and machine learning techniques. All analyses were performed on four datasets organized into two groups: those who completed the regular version of the reading comprehension measure (BALA), and those who completed the modified version. The regular version of BALA was the default version for participants in Grades 5-6. The modified version was linguistically simplified and it was administered to all Grade 4 students as well as students in Grades 5 and 6 whose teachers recommended it based on their having reading difficulties or being at earlier stages of English language acquisition.

There were two elicitation methods: text and oral. The oral-elicited data were derived from participants' spoken responses to two types of tasks: picture description (two tasks) and story retell (two tasks), which were then transcribed by a team of graduate students. The text-elicited data was derived from participants' responses to a writing prompt about social media and from their written responses to open-ended questions about narrative and nonfiction texts. Each language sample was processed through the COVFEFE NLP pipeline to extract lexical and syntactic language features. To reduce task effects, and to consolidate the data for ML, the NLP-derived features were aggregated across tasks, by taking the mean of each feature across tasks within elicitation methods, culminating with four datasets: oral- and text-elicited, for participants who completed the regular and modified versions of the reading comprehension assessment.

5.1 Supervised methods

Hagtvet (2003) administered the same language production tasks in aural and written form and found it remarkable that skills across oral and written domains were highly correlated. The present study may be the first to use fine-grained NLP-derived features that are identical in both speech- and writing-elicited data, thus aiding comparison across domains in terms of modeling their associations with reading comprehension.

First, supervised ML methods were employed to examine whether the 260 NLP-derived syntactic and lexical features could predict reading comprehension. Eight models were compared to a mean baseline for the best performance, and the models with the lowest error for each of the four datasets were selected. For both the regular and modified text-based elicitation datasets, random forest was the best algorithm. Using cross-validation, the relative error reduction for text/regular was 12.75 percent but only 2.4 percent for the text/modified. For oral/regular, the support vector machine model had the best performance (a relative reduction in error of 6.07 percent), while the gradient boosting algorithm proved the best for the oral/modified data, relatively reducing the error by 21.17 percent. No clear pattern exists for which reading comprehension version or which language domain among the four models was best in terms of relative error reduction; the oral/modified had the best error reduction, followed by text/regular, then oral/regular, and lastly text/modified. In other words, the ML algorithms were more successful with some datasets than others, but not in a readily interpretable way. (The small sample sizes contribute to this result, with the test portions of datasets ranging in size from only 13 to 21 participants.) Nonetheless, the results extend the research base around the relationship between oral language and reading comprehension by finding that substantial variance in the latter can be modeled from productive (rather than receptive, binary scored [correct/incorrect]) lexical and syntactic features extracted through natural language processing from children's speech and writing.

In the supervised models for RQ1, relative error reduction was based on a cross-validation approach where the data were partitioned into two parts, independently, 200 times, with an 80/20 train/test split. The algorithm learned the training data — that is, it learned the relationship between reading comprehension and the lexical and syntactic speech features in the training data — 200 different times, each time applying that learning to predict the reading comprehension score of the test data. Then, the results were averaged to attain the average mean absolute error between predicted and actual values in the test data. Therefore, the predicted values reported here are quite literally predictive — in the sense that the algorithm is designed to "predict" the reading scores of new, unseen data. Thus, when comparing results to traditional regression studies, it is critical to remember that this is not simply "variance explained" being reported, e.g., the amount of variance in the reading comprehension score explained by the linguistic features. Although exploratory, the goal of the analysis is for generalizability, that is, to be able to accurately predict

the reading scores of previous unseen linguistic feature (test) data, based on what was learned about that relationship in the training data.

Each algorithm's performance on a single test-train split could be greater or less than the cross-validated result, suggesting that the small sample sizes play a role in the instability of the modelling: high-leverage and outlying observations can strongly influence the model's predictive power. In future studies with larger datasets, the models may be more successful. Nonetheless, even with the small sample, the predictive power and amount of variance explained by these ML models do compete with existing studies: the range of variance explained in a single train-test instance (R^2) was 18.5 – 36.6% across the four models. Table 33 provides a summary of published studies that modelled the relationship between reading comprehension and either syntax, or grammar, or both. It is provided here for reference when comparing the outcomes of RQ1 to published research. To improve comparability with the results of the present study, Table 33 summarizes studies that model reading comprehension as a continuous variable. Some studies reviewed in Chapter 2 are not included in this table because they used a reading comprehension score to group participants and then used ANOVA or factorial design (e.g., Nation & Snowling, 2000). Although these contribute a great deal to the field, their design (group comparisons) prevents as facile comparisons as is possible with studies using continuous reading comprehension outcome variables.

The studies summarized in Table 33 suggest a range of predictive values for lexical and syntactic features in the prediction of reading comprehension. No study exactly matches the present study in its design, so these are provided for general comparison purposes only. When cross validation was used, the improvement from a mean baseline ranged from 2.40 to 21.17%. When a single train/test split was created for each of the four models, the variance explained (R^2) was between 19-31%. Both are within the range of the studies summarized in Table 33. Poulsen and Gravgaard (2016) researched Grade 5 students, similar to the present study) and found that after decoding and fluency were held constant, an addition 13% of variance in reading comprehension could be attributable to vocabulary, and 12% to syntactic knowledge. While these cannot be assumed to be independent, the sum of these (25%) is similar to (slightly above) the best predictive model, which was 21%. Again, sample sizes must be taken into consideration.

An important difference between the present study and the studies cited in Table 33, however, is that the present study did not explicitly evaluate participants' vocabulary or grammar using scored measures, but instead used a descriptive dataset of NLP-derived linguistic features extracted from participants' spoken and written language output, rather than selected-response or other traditionally scored measures. No other language-related factors were held constant in the present modelling (i.e., language comprehension, verbal working memory), unlike studies such as Gottardo et al. (1996) and Tunmer and Chapman (2012) which do hold such constructs constant and report lower variance explained by either grammar or lexis.

Table 33

Summary of studies investigating vocabulary and grammar as predictors of reading comprehension

Source	**Participants**	**Result**
Catts et al. (1999)	Longitudinal study, 604 participants in kinder-Grade 2 (L1)	Kindergarten grammar composite correlated with Grade 2 RC at .66. Kindergarten vocabulary composite correlated with Grade 2 RC at .50. Kindergarten oral language composite had .48 partial correlation with Grade 2 RC.
Poulsen & Gravgaard (2016)	80 Grade 5 participants (Danish L1)	After a decoding and fluency composite score was held constant, vocabulary explained an additional 13% variance in RC and comprehension of syntactically basic and difficult sentences explained 12% RC
Deacon & Kieffer (2018)	Longitudinal study, 100 participants in Grades 3-4 (L1)	Using path modelling to predict Grade 4 RC, Grade 3 standardized regression paths were .21 for syntactic awareness and .05 for vocabulary, after controlling for Grade 3 autoregressive RC.
Brimo et al. (2017)	179 Grade 9 and 10 students (L1)	Using path modeling, syntactic awareness indirectly influences RC (significant indirect effect on RC of .06) through syntactic knowledge (significant direct effect on RC of .15). Vocabulary also has significant direct effect of .59.
Demont and Gombert (1996)	Longitudinal study, 23 participants ages 5-8 (French L1 study)	Grade 3 measurement of correction of asemantic and agrammatical sentences predicted 24% of variance in Grade 3 RC, after holding nonverbal intelligence, IQ, and vocabulary constant.
Siegel (2008)	1,238 Grade 6 participants (309 learning EAL; L2)	Syntax predicted 9-19% of variance in a range of reading comprehension when modeled with morphological awareness.
Gottardo, Stanovich, and Siegel (1996)	112 Grade 3 participants (L1)	Holding constant verbal working memory and phonological sensitivity, syntactic processing accounted for 1.3 to 1.5% of unique variance in predicting RC
Cain (2007)	196 participants ages 7-8 and 9-10 (L1)	For younger group, word-order correction predicted 16% of RC variance; 15% for the older group. However, if entered as a second step after vocabulary, memory, and receptive grammatical knowledge, the contribution of word-order correction was <2% for both groups. In a separate analysis, grammatical correction predicted 8% of RC variance

		for the older group, but when entered as second step after vocabulary, memory, and syntactic awareness, it accounted for <1% of variance.
Ouellette and Beers (2010),	67 Grade 1 participants and 56 Grade 6 participants (L1)	Vocabulary breadth predicted reading comprehension in the older group (but not the younger group) with a .15 change in R2 when entered into a hierarchical regression after phonological awareness, decoding, irregular word recognition and language comprehension.
Ricketts, Nation, and Bishop (2007)	81 participants ages 8-10 (L1)	Vocabulary predicts an additional 18% of variance in RC after age, IQ, decoding, regular word reading, and exception word reading were entered into a hierarchical regression.
Ouellette (2006)	60 typically developing Grade 4 students (L1)	Vocabulary breadth, as measured by expressive (picture-naming) and receptive (picture-pointing) tasks, and vocabulary depth (word-defining and synonym-identification tasks) predicted RC, with vocabulary depth explaining an additional 8% unique variance when entered last in regression equation after age, nonverbal IQ, visual word recognition, decoding, and vocabulary breadth (which itself contributed 7%).
Tannenbaum et al. (2006)	203 Grade 3 students (L1)	Lexical depth, breadth, and fluency predicted a total of 50% of the variance in reading comprehension, with depth and fluency forming a factor that contributed 19% unique variance, breadth contributing 2% unique variance, and 29% of variance being common to the two factors.
Cain & Oakhill (2014)	83 participants ages 10-11 (L1)	After holding age and decoding accuracy constant, vocabulary breadth depth predicted 6% of variance in ability to make local cohesion inferences, but depth was not a significant predictor. In predicting global cohesion inferences, vocabulary breadth predicted 17% and depth predicted 8% additional variance.
Tunmer and Chapman (2012)	122 Grade 3 participants (L1)	Entered as the final step in a hierarchical regression, vocabulary predicted 2% additional variance after age, language comprehension, word recognition, and letter-sound knowledge.
Verhoeven, Leeuwe, and Vermeer (2011)	Longitudinal study, 2,790 participants grades 1 to 6 (Dutch study included both L1 and L2)	Grade 2 beginning-of-year basic vocabulary predicts Grade 2 end-of-year RC, which in turn predicts Grade 3 advanced vocabulary, which in turn predicts Grade 4 RC. Grade 5 advanced vocabulary also predicts Grade 6 RC, demonstrating some reciprocity, but as students advance through school the direction is from vocabulary to RC.
Proctor, Silverman, Harring, & Montecillo, 2012	294 participants in grades 2-4, 44% were bilingual (English-Spanish), half learning EAL	For all participants, English vocabulary breadth and depth significant predicted RC at the initial timepoint (significant standardized estimates were .19 for breadth, .18 for semantics, and .13 for morphology). No vocabulary measure predicted growth rate. Bilingual learners' vocabulary in the L1 (Spanish) did not predict L2 reading comprehension.
Gottardo et al. (2018)	52 bilingual (Spanish-English) participants, ages 8-13	Vocabulary, morphology and syntax explained 67% of the variance in RC. The three factors in common explained 41% of the variance, with 9.83% uniquely contributed by vocabulary, 9.02% by syntax, and 4.38% by morphological awareness.
Foorman et al., 2015	1,792 participants in grades 4-10 (primarily L1)	A general oral language factor measured by syntax and vocabulary predicted RC at all grade levels with standardized coefficient estimates of .72 to .96. The specific vocabulary factor was a significant predictor only in Grade 7 (.18 standardized coefficient), and the specific syntax factor was only significant in Grade 4 (.50 standardized coefficient).

Note: L1=study focused on L1 acquisition; L2=study focused on L2 or L1 & L2 acquisition; RC=reading comprehension

The second part of RQ1 sought to understand which lexical and syntactic features were important predictors of reading comprehension in the four best ML models. Random permutation was performed using a cross-validation approach with the ELI5 package in Python. The best models (as described above) were run, using 200 train/test partitions for cross-validation; for each cross-validation run, one variable was randomly permutated, or scrambled. This process is repeated for each variable in the dataset. The average increase in the mean absolute error is the metric used to gauge each variable's contribution to the model. The top twenty feature predictors for each of the four models (oral/text by regular/modified) were reported. Broadly speaking, grammatical constituents were more commonly found among top predictors in the regular versus the modified datasets (Figure 17). This leads to the conclusion that grammar-related features may predict reading comprehension skills for older or more skilled readers better than for younger or less skilled readers. (However, grammatical complexity [rather than constituents] was not well represented in any of the supervised models, with only two or fewer features present in the top predictors of each.) In contrast to Catts, Fey, Tomblin, and Zhang (2002) who found that grammar was the language domain most closely associated with reading comprehension outcomes in both Grade 2 and Grade 4, but there was no difference between grades, the present findings suggest that syntactic features may become more predictive of reading comprehension for older or more skilled readers. The finding that grammatical features were more important in the regular versus modified data models also relates to Geva and Farnia's (2012) finding that syntactic awareness was a significant predictor of reading comprehension in Grade 5 only for EAL learners, and not for students who were non-EAL learners. In the present study, participants learning EAL were present in both modified and regular versions of the data, so no firm conclusion can be drawn with regard to Geva and Farnia's (2012) finding. Future research could parse the EAL and non-EAL learners into separate groups to examine how NLP-derived syntactic variables predict reading comprehension for both groups.

Grammatical constituents were more commonly found in the top predictors of the oral-elicited datasets than the text-elicited datasets for both modified and regular BALA (Figure 17), suggesting the nature of grammar in speech may a stronger predictor than grammar in writing. Studies cited above that measured grammar – whether through grammatical or word-order correction, syntactic knowledge, or grammaticality judgment tasks – largely use oral measures (rather than written measures) (i.e., Cain, 2007; Demont & Gombert, 1996; Gottardo et al., 1996;

Gottardo et al., 2018). In addition, those measures are most often selected, rather than constructed, responses. Thus, a ready comparison cannot be made with the results of the present study that concern the relatively higher importance of grammatical features in the oral data versus the written data. The nature of grammar in spontaneous speech and in writing, and its relationship with other literacy skills, remain ripe areas for research.

Like the grammatical features, the pattern for vocabulary richness and range variables had greater representation in the top features of the regular data than the modified data. This could relate to the nature of some of the subjective vocabulary metrics, which were elicited from adults. Unlike the grammatical features, though, the top features of the text-elicited data had greater representation of vocabulary richness and range than oral-elicited data. This suggests the sophistication of vocabulary in writing may be more strongly associated with reading comprehension than sophistication of vocabulary in speech. However, this finding could relate to task effects, as the vocabulary in the oral elicitation tasks may not have provided ample opportunities to use sophisticated vocabulary.

With regard to vocabulary sentiment and affect, the modified datasets (especially oral-modified) were more likely than the regular datasets to have vocabulary sentiment and affect variables among the top predictors. Vocabulary specificity, similarity, and ambiguity was twice as common in the top predictors of the text/modified model than in other models. These are discussed further below. Overall, these patterns suggest that for less skilled and younger readers, the affect and sentiment in their productive language is more predictive of their success in reading than the vocabulary or grammatical constituents they use; however, this finding may not be generalizable because the nature of the sentiment and affect predicting reading comprehension in these data may not be equivalent to that which would be predictive in other tasks. This finding requires further investigation.

5.2 Interactions with demographic variables

The second research question investigated interactions between two demographic variables (years in Canada and multilingual proficiency) and the top lexical and syntactic features in predicting reading comprehension outcomes. A pairwise permutation process was used to identify pairings of two demographic variables and the top five linguistic feature predictor variables as candidates for a more confirmatory interaction approach using multiple regression.

This process is akin to differential item functioning analyses that hold the outcome constant and examine whether scored assessment items favor a certain demographic group. Here, the demographic variables are included in the model to determine if the predictive slope (between a given linguistic feature and the reading comprehension score) differs across groups. This is an important area for research, because generalizations about the functioning of NLP-derived variables in predicting any construct must be proven to be fair and reliable across demographic groups (Madnani, Loukina, von Davier, Burstein, & Cahill, 2017).

There were significant interactions in both modified models: in the oral/modified data, the use of a complex verb phrase significantly interacted with years in Canada in predicting reading comprehension, and, in the text/modified data, a significant interaction occurred between word imageability and years in Canada. Kim and Jang (2009) found that differential functioning of vocabulary and grammar items on a literacy assessment can occur for students learning EAL; specifically, they found that vocabulary items tended to favor non-EAL students and grammatical items tended to favor EAL students. The multilingual proficiency variable did not have significant interactions with any of the top predictors, suggesting that in this sample children's multilingualism is not impacting how well a given NLP-derived syntactic or lexical variable predicts reading comprehension.

In this case, no conclusion can be drawn about whether one group is favored, but instead the focus is on the relationship between these linguistic features and reading comprehension. It is possible that the verb phrase interaction relates to the verb form in question: a verb phrase consisting of "to" and a verb phrase. For a child still acquiring English, this somewhat complex verb form may not be mastered, and as seen here, its use is closely related to broader literacy skill. For participants who have been in Canada for 10 or more years, the use of this verb phrase may be ubiquitous, and therefore does not predict a broader literacy skill such as reading comprehension. As for interaction with imageability in the text/modified data, the lack of negative relationship for the participants who had lived in Canada for 10 or more years is difficult to explain, since vocabulary metrics should be fairly universal. It would be fruitful to examine these further to determine if patterns exist with regard to differential functioning of syntactic or lexical variables across demographic groups.

For both significant interactions, the group that had been in Canada longer had flatter slopes for both interaction models, indicating that the language features were less predictive for this group. The reason for the weaker relationship for students who had been in Canada longer is not answerable with the current data. Follow-up think-aloud protocol that allows participants to reflect on their performance, or perhaps the inclusion of standardized grammar and vocabulary measures could prove fruitful in teasing apart the reasons for these interactions. It is interesting to note that no significant interactions were produced using the text/regular data. This could indicate that the validity of these features in language and literacy research is potentially threatened when applied to the language of less skilled or younger learners (the modified datasets). Further research is necessary to understand the nature and cause of these interactions.

5.3 Unsupervised ML methods

Finally, an unsupervised ML approach was executed to understand the underlying structure of the linguistic feature datasets, using factor analysis in Scikit-learn. The factor analysis results explained 14-28 percent of the variance in each dataset. Overall, the pattern of feature loadings showed that the construct of grammatical complexity was more predominant in the regular than modified data, especially for the oral-elicited data. This suggests younger learners and those who are acquiring EAL may not use sufficiently complex grammar for it to be a strong factor; however, the *length* of clauses and sentences did load onto a factor in both text/modified and oral/modified datasets.

Vocabulary richness (lexical diversity) correlated positively with grammatical complexity for the regular dataset, but there was no evidence of such a relationship in the modified data, which again may relate to developmental differences and the overall lack of grammatical complexity factor loadings in the modified data. These findings suggest the relationship between vocabulary richness and grammatical complexity may change over the course of language development. Vocabulary range did not correlate strongly with either of these, except in the oral/modified data, where vocabulary richness and range loaded in the same direction. The vocabulary specificity, similarity, and ambiguity features were present in several factors but are somewhat difficult to interpret. Sentiment and affect features were only present in the oral/modified data and very slightly in the oral/regular data.

In regression analyses, a factor composed of grammatical complexity and vocabulary richness positively predicted reading comprehension in the oral/regular dataset. In the text-regular data, three factors predicted of reading comprehension: one comprised of grammatical complexity and vocabulary richness, one of clause length and total length, and a third of vocabulary range and word specificity. These align with the literature base that finds vocabulary diversity, use of grammatically complex phrases, and vocabulary range being positively associated with reading comprehension (Deacon & Kieffer; 2018; Foorman et al., 2015; Gottardo et al., 2018; Ouellette, 2006; Ouellette & Beers, 2010; Poulsen & Gravgaard, 2016). In the text/modified data, a word similarity (mean and variation) factor was a significant negative predictor of reading. This suggests highly repetitive writing is negatively associated with reading outcomes. This negative relationship does not hold true in all cases, however, as word similarity appeared to support writing cohesion in other data.

The next two sections describe the findings specifically relating to grammar and vocabulary.

5.4 Grammar

Overall, grammatical complexity had a stronger presence in the unsupervised factor analysis approach (RQ3) than in the supervised ML approach (RQ1). In the unsupervised models, grammatical complexity was a strong organizing construct – that is, a construct around which the factors agglomerated – for the regular datasets. In the supervised models, only five total grammatical complexity features were among the four models' top twenty predictors – when clause and sentence length (rather than complexity) features are omitted. These five are complex nominals per clause (negative predictor in oral/regular models), dependent clauses per clause and length of prepositional phrases (positive predictors in oral/modified models), and number of clauses and T-units (negative predictors in text/regular models). (This last result is not certain, as it relies on inferring that a lower number of clauses or T-units correlates with each clause or T-unit being more complex. This inverse correlation was noted in several analyses; still, it should not be interpreted as completely reliable.) Yet in the unsupervised approach, two of four factors in both regular datasets, and one of five factors in the text/modified (but none of the three oral/modified factors) were oriented around grammatical complexity. Of these, one grammatical complexity factor predicted reading comprehension in the oral/regular data, and two in the text/regular data did so.

A takeaway from this comparison is that grammatical complexity is an important organizing construct of productive language – as is well known – and the factors that organized around grammatical complexity do predict reading comprehension. However, as an individual feature, grammatical complexity may not directly predict reading comprehension when many other fine-grained lexical syntactic variables are present. This paradoxical finding aligns with existing research because the specific role of syntactic ability is still debated. While some scholars have found syntactic ability to be a strong predictor of reading comprehension (e.g., Demont & Gombert, 1996; Deacon & Kieffer, 2018) others question the role that syntactic complexity alone plays in reading comprehension. For example, Gottardo et al. (1996) held verbal working memory and phonological sensitivity constant and found that syntactic processing accounted for only 1.3 to 1.5% of unique variance in reading comprehension, while Cain (2007) had similar results when adding syntactic knowledge as the last step of a hierarchical regression. Catts et al. (2002) conclude that although grammar was an important predictor of reading comprehension in their models, this does not necessarily mean grammatical processing itself contributes to reading struggles, or if less well-developed grammatical skills are an index or marker of broader language challenges. This caveat has also been discussed in Nation and Snowling (2000) and Bowey (1994). The same qualification must be applied to the present study, especially since other variables were not held constant (e.g., listening comprehension) in the models. Further research is needed to understand why factors organized around grammatical complexity were more likely to be strong predictors of reading comprehension than the individual grammatical complexity features. Since grammatical constituents, rather than grammatical complexity, were prevalent predictors in the supervised models, it is possible that a bifactor or other hierarchical model may better fit these data (e.g., with grammatical complexity as a level above grammatical constituents).

However, all grammatical constituents may not be related to grammatical complexity. For example, the supervised models appeared to associate noun usage with lower reading outcomes, and verbs with greater reading outcomes; the latter was especially true for the oral tasks, where narratives including verbs seemed to preclude a "listing" style of description. Although nominalization is a key to academic language (Fang et al., 2006), these findings better align with Crossley et al.'s (2014) finding that noun use is not a key element of human raters' evaluation of writing. In fact, here, noun use is associated with lower reading outcomes. As Gentner (1982,

2006) noted, verbs are more difficult to learn than nouns are, and considering the age group under focus here, this may indeed be an underlying cause for this finding.

5.5 Vocabulary

In general, vocabulary range metrics operated exactly as anticipated: age of acquisition and word length positively predicted reading outcomes, while frequency, imageability, and familiarity were negative predictors. This is not surprising given Gilhooly and Logie's (1980) finding that familiarity, frequency, and imageability are positively correlated, and they all correlate negatively with age of acquisition. However, other vocabulary features are more nuanced or opaque. Vocabulary specificity and ambiguity, as operationalized through the WordNet corpus, did not offer a tidy interpretation. Average ambiguity, which is expected to positively predict the reading outcome (Casas et al., 2018), was indeed a positive predictor of reading comprehension, but only in the text/modified data (specifically the ambiguity of nouns). Given that the research on ambiguity and language development finds that children's use of ambiguous words ceases to differentiate from adults' use after age 5 or so, this finding makes some intuitive sense, since the participants who completed the modified assessment were either younger or less proficient in English or literacy. *Functions* of ambiguity were strong positive predictors, though, which although was not expected, does make sense intuitively: if only ambiguous words are used, then less higher-level vocabulary can be used, since ambiguous words usually are higher frequency words. Thus, the function of variation in ambiguity, which was a positive predictor for oral/modified and text/regular, merits further inquiry.

Word similarity, which measures the similarity of words across a given passage, had positive predictions for text/regular, but negative predictions for text/modified. An example given in that section's discussion illustrated how word similarity could benefit writing cohesion, but overuse of similar or identical words was simply repetitive; participants' developmental or language acquisition stage could be impacting this result. In the unsupervised data analysis of the text/modified data, a factor composed entirely of means and standard deviations of word similarity was a significant negative predictor, aligning with the supervised model's results for this dataset. This finding agrees with Crossley et al. (2011) who found that content word overlap in writing tended to diminish from younger to older participants (they were adolescents and young adults, though, not children). The developmental trajectory may not be linear and merits

further exploration. Specificity of verbs positively predicted reading outcomes for the text/modified data, as might be expected, even though specificity is not directly correlated with vocabulary range, as discussed above. For the two oral datasets, variation in specificity was an important predictor, but it operated positively in oral/regular and negatively in oral/modified. Again, the question of variation in these features would benefit from additional research. On the whole, the text/modified had eight vocabulary specificity, similarity, and ambiguity features among the top 20 predictors in the supervised model, which was more than twice of any other dataset. Further investigation will be necessary to understand whether this is due to a lack of developmentally appropriate vocabulary corpora in the COVFEFE pipeline, or if these features are indeed as important as the current results suggest.

Vocabulary sentiment and affect variables were present in all supervised models, with greater representation in the modified models. This may relate, again, to developmental stages and students' self-regulation. Including other data sources such as students' self-reports around self-regulation could address this question. The sentiment and affect features that were important predictors appear to not be generalizable across tasks, for example, the oral elicitation tasks included some quasi-dramatic events that required use of negative-affect vocabulary to retell accurately. Hence, the positive predictive value of negative affect makes sense there. As Harris (2018) notes, sentiment analysis tools trained on natural language lack generalizability as they are highly dependent on the corpus on which they were trained. Comparing the results of different sentiment and affect corpora could be a fruitful line of study.

The other affect and sentiment features are difficult to interpret: does the negative predictive value of arousal in oral/modified indicate that participants were uncontrolled in their excitement? In their original development of the corpus, Warriner et al. (2013) found that younger individuals provided higher ratings on arousal and valence. But in the text/modified version, mean arousal is a positive predictor of reading comprehension – does this have to do with text and speech differences, where arousal in written measures means engagement? Further research is necessary.

Of note was the absence of sentiment and affect features among the factor analyses, except the oral/modified data (which aligns with the supervised model of that data) and very slightly in the oral/regular data. This is the opposite scenario from the grammatical complexity paradox, where

few grammatical complexity items were top predictors in the supervised models, but they were important organizing constructs in the unsupervised models. This suggests sentiment and affect variables can predict reading comprehension (although this may be task-specific and not generalizable across tasks), but sentiment and affect features may not be strong enough constructs in general language production for unsupervised factors to organize around them. In the unsupervised modeling, the factor in the oral/modified data that was closest to achieving statistical significance in predicting reading comprehension was the one with strong loadings from affect and sentiment variables, which does indeed align with the supervised model for that dataset – but the other datasets did not show this pattern. Again, further investigation of the role of affect and sentiment in reading ability, for readers of different skill level and/or ages, would be a fruitful line of further research.

Vocabulary richness, also known as lexical diversity, had greater presence in the unsupervised than supervised models, similar to the scenario for grammatical complexity. Only two vocabulary richness features were among the top predictors in the supervised models, and both were for the regular datasets. In oral/regular, Honoré's statistic (a function of how many words are used only one time in a language sample) was a positive predictor, while in text/regular, the moving average type-token ratio (30-word window) also positively predicted reading comprehension. In the unsupervised models, moving average type-token ratios were present in 2 of 4 factors for each regular model, 2 of 3 factors in oral/modified, and 1 of 5 factors for text/modified. In both regular datasets, factors with moving average type-token ratios significantly and positively predicted reading outcomes. In oral/regular, the moving average type-token ratio features loaded in the same direction as sentence length and total length. In text/regular, moving average type-token ratios loaded with grammatical complexity for one factor, and with sentence length and total length for another. It appears the moving average type-token ratio is an important construct in all four datasets, since these features are present in factors in all four datasets, but these factors were predictive of reading comprehension only in the regular datasets. This is an unexpected finding. Again, it may relate to differences in language development or maturity across the modified and regular datasets, but this will have to be confirmed with future research.

In all factors on which moving average type-token ratio features loaded, they loaded with features that positively predict reading outcomes. The standard type-token ratio, which calculates

vocabulary richness across the entire language sample, loaded with lexical and syntactic features representing lower language and literacy abilities, while the moving averages consistently loaded with features representing higher abilities. Also of note is that in factors where both the standard and moving average type-token ratios loaded, they consistently loaded inversely to one another. This is evidence in support of the critiques of the standard type-token ratio (e.g., Chipere et al., 2001; Wood et al., 2019; Covington & McFall, 2010; Kapantzoglou et al., 2019).

5.6 Comparisons across the four datasets

The supervised and unsupervised models developed for the four datasets share elements in common but all have unique results. The oral/modified data had the greatest relative error reduction in the supervised model, yet paradoxically, none of the three factors of the oral/modified dataset that were produced during unsupervised modelling were statistically predictive of reading comprehension. In addition, the top 20 predictors in the supervised modelling of the oral/modified data were more difficult to interpret, with vocabulary sentiment and affect features (which are challenging to understand in light of the literature base) as 6 of the top 20, and several of the other top predictors in this model also resisting easy interpretation. It is possible that this is due to the corpora and processes used for NLP in the COVFEFE pipeline; that they were created for and normed on adults may make their application to a younger or less linguistically developed population less appropriate and more difficult to interpret. Some features that are well known to be associated with higher language and literacy proficiency – vocabulary richness and grammatical complexity – were found to be strong predictors in the regular data. On the other hand, sentiment and affect were more likely to be strong predictors in the modified data. More research is necessary to understand if this is an artifact of the NLP feature extraction methods, or if it is indeed related to developmental or language and literacy proficiency differences.

The use of coordinating conjunctions was negatively predictive across all four datasets: the use of “and” to connect ideas created long, but not complex, sentences. This study demonstrates there is a substantive difference between the two, and aligns with work by Carretti et al. (2016) that found less-skilled comprehenders tended to use additive connecting words (“and”) rather than causal connecting words (“because”). However, grammatical complexity was a more important predictor for the regular data than the modified data, suggesting for the younger/less

skilled readers comprising the modified data, the relationship between sophisticated syntax use and reading comprehension is not as strong as for the regular-BALA dataset.

In addition, pronoun use was a negative predictor in both text-based datasets, which suggests that this may represent a difference between written and spoken language; indeed, some participants addressed their writing to "you", which lacks formality. While many grammar and vocabulary features transcended oral and written elicitation methods, pronoun use was only a significant negative predictor in the text-elicited data. It would be interesting to follow up this study with another study similar to Hagtvet (2003) where the exact same language production tasks are administered in both aural and written form, to compare the NLP-derived features for each and see if the results obtained are similar to Hagtvet's, where she found high correlations between all measured constructs elicited in text and oral form.

Vocabulary range, or the use of sophisticated vocabulary, was a strong predictor across all datasets. Grammatical complexity and vocabulary richness correlated in both regular datasets (but not modified), while vocabulary richness and range were only correlated in the oral/modified data. Interestingly, vocabulary range emerged as operatively somewhat more independently than vocabulary richness and grammatical complexity; however, the independence (or correlation) of these constructs appears to differ depending on the developmental stage under consideration.

6 Limitations and conclusion

This study aimed to explore the potential benefits and limits of NLP tools and ML analytic techniques in language and literacy research. This work was inspired by, and aims to contribute to, the extensive literature describing the cognitive and linguistic constructs that are associated with successful reading comprehension. The study is exploratory for many reasons: first, factors such as working memory, decoding, and age were not held constant; the unsupervised modeling approaches were exploratory (rather than confirmatory) factor analysis; beyond that, the corpora used in the NLP are "out-of-the-box" and have not been validated on children. Nonetheless, the study aimed to demonstrate what may be possible with these tools, and to hopefully encourage further research that utilizes them.

In terms of limitations, a possibility exists that the open-ended reading comprehension questions used in these analyses may not be completely independent from the reading comprehension outcome measure. However, several factors combine to lessen the potential for concern. First, none of the open-ended responses were evaluated for their content or "correctness". In fact, three of the four types of open reading responses utilized here – make open-ended predictions about what will be read, describe interest in the text, generate three questions about what was read – do not have correct or incorrect answers. The final, high level comprehension question is quite open-ended but does require text evidence; however, again, it is not evaluated specifically for content correctness in the present study. While a lesser understanding of the text may result in a less sophisticated response to these four questions, not fully understanding the text after reading does not preclude creating a lexically and syntactically sophisticated response.

With regard to analysis design, in studies of children examining the relationship between multiple intraindividual variables, spuriousness may exist due to the potential confounding factor of age (Walford, Tucker, & Viswanathan, 2010). Age could feasibly explain, in part, the relationship between lexicosyntactic variables and reading comprehension. Indeed, some extant studies examining the predictive relationship between different language and/or literacy components, often hold age constant. However, in the current study, age was not held constant (it was not "partialed out"). For one, the research team did not have access to student records or parents/guardians to verify birthdates. However, within each sample (BALA-modified and BALA-regular) the age ranges were fairly truncated, spanning only one to two years with just a few exceptions.

Nonetheless, concern about this possible limitation can be alleviated by two points. First, peer-reviewed research studies analyzing the relationship between different components of language and literacy do not uniformly consider age when using regression-based analyses (e.g., Babayiğit, 2015; Brimo et al., 2017; Lesaux et al., 2007). Even a meta-analysis looking at the relationships between reading and writing (Graham & Hebert, 2011) did not include age as a factor of interest. These studies establish a clear predictive link between different literacy constructs, despite not including age as a control variable.

One study of the relationship between reading and vocabulary cited in this work, by Cain and Oakhill (2014), did include age in their hierarchical linear regression. When entered alone as the

first step in the hierarchical regression (p. 21), age accounted for less than 1% of variance explained in reading comprehension (inference) scores. The participants in Cain and Oakhill's study were 10-11 years old, roughly mirroring the present study (approximately 9 to 11 years old). Therefore, the potential limitation of not including age in the present study is minimized by 1) the corpus of existing research that has established the intra-individual relationship between literacy constructs, without including age; and 2) the published study by Cain and Oakhill (2014), which used a sample similar to that of the present study and found less than 1% variance in inferential reading comprehension explained by age. Nevertheless, an important future step would be to structure the research design in such a way as to hold more constructs (such as age and working memory) constant, such as studies by Oakhill, Cain, and colleagues (e.g., Cain, 2007; Cain & Oakhill, 2014), and investigate the ML models' predictive power when such a set of covariates is included.

Literacy research that includes socioeconomic status has been very useful in understanding literacy development, and, indeed, academic achievement broadly speaking. The positive association between SES and academic proficiency has been found across cultures and geopolitical boundaries (Chen, Lee, & Stevenson, 1996). In his meta-analysis, Sirin (2005) indicated that family-level SES has a medium degree of association with academic performance in K–12 settings and that school-level SES has a large degree of association; the overall finding is that family SES impacts achievement in multiple ways, including determining the school students attend, assuring the resources and social capital necessary for school success, and maintaining a relationship between families and school. As the findings around SES are largely conclusive, I focused on variables that are critically relevant to this study: reading comprehension, myriad lexicosyntactic variables, and variables relating to multilingual proficiencies and immigration background. Future research may consider incorporating other demographic information, such as rural or urban school setting, ethnicity or race, and SES, to examine relations with the variables used in the present study.

It is hoped that the study is read as a somewhat cautionary tale. The NLP tools are quite robust, and they offer very fine-grained information about productive language. Yet in their robustness, they are also very sensitive to the input, particularly task characteristics. The datasets analyzed in this study had been aggregated (averaged) within task types to address data sparseness. Some high-leverage datapoints were investigated, and they turned out to be biased due to missing data

for one or more tasks. For example, if the participant only completed the tasks that tended to elicit shorter responses, then their results could be biased in a variety of ways. Research using these tools would benefit from ensuring that tasks do not elicit construct-irrelevant variance, that the length of all language samples are within a certain range, and that data are as complete as possible. Wood et al. (2018) found that expressive vocabulary measures utilizing shorter samples (i.e., 50 words or less) are less valid than longer samples, so erring on the side of longer rather than shorter would be beneficial. While ML tends to have an aura around it – that it can solve any statistical problem – and it *is* a quite powerful set of statistics – the old adage of GIGO ("garbage in, garbage out") is just as apt here as in traditional statistics.

Typically, research using NLP in the field of language and literacy education and assessment tries to approximate human raters (e.g., Black et al., 2009; Somasundaran et al., 2015). While that is quite important work, the goal of this study was to explore what NLP and ML themselves may have to offer the field in terms of theory. The importance of lexicosyntactic features was not examined in terms of a comparison with human raters, but instead employed machine learning techniques to ascertain which features were associated with strong reading comprehension. Are the features that emerged as important predictors of reading comprehension similar to those of human raters, as reported in the literature? As expected, the findings are somewhat inconclusive, although there is little doubt that some elements can readily support current theory and practice. The functioning of grammatical complexity and vocabulary richness and range features clearly align with the established literature around such constructs (although the differentiation in factor analysis between vocabulary range and the other two mentioned here was unexpected). The role of more novel features – grammatical constituents, vocabulary specificity, similarity, and ambiguity, and affect and sentiment – requires more thorough analysis. It is hoped that future studies can home in on some of these features to determine if they are useful and can contribute to the theory and practice around language and literacy learning and assessment. Further, it is hoped that studies using NLP can investigate changes in these constructs over the developmental span, in terms of both L1 and L2 language use. As always, it is essential that the models are valid across demographic groups.

The contemporary conceptualization of validity focuses on inferences that are made about learners based on their assessment scores, with the primary question being: are the inferences appropriate? The present study provides an opportunity to consider a very large number of

descriptive features of learners' language, and examines the relationship of those features to reading comprehension. A causal relationship cannot be ascertained from the present study, that is, whether reading comprehension supports the development of sophisticated lexis and syntax, or vice versa. This would be a ripe area for future research. Nonetheless, this study can already inform assessment theory and practice by offering an extremely fine-grained assessment of spoken and written language.

As a qualifier, however, when something is called "assessment," it means there must be an inference or action taken. In a nutshell, there must be some consequence or form of feedback. Future research may consider using these robust NLP and ML techniques to predict human-developed writing scores, and to consider whether the emerging features may be valuable for teaching and learning. For example, should young writers consider how to attend to affect and emotion in their informational writing, which is not currently typically a focus in this area of writing instruction? One vast area of potential for NLP and ML is to uncover "hidden" patterns in language that humans may not recognize explicitly, such as the emotions and affect discussed in this study. This powerful pattern recognition ability can also support deep analyses of bias in writing scoring, which would be another welcome area for future research.

There is great potential for the use of supervised machine learning for both research and day-to-day language and literacy learning and assessment (for example, to provide just-in-time assessment results to teachers and learners). However, the sheer number of features make interpretation a challenge, and further research is needed to understand the best way to consolidate features into meaningful constructs for both research purposes and learner feedback. The factor analyses in the third set of research questions attempted to do this, with mixed results in terms of interpretability. Nevertheless, a critical takeaway from this study is that both spoken and written language can predict children's reading comprehension, aligning with the body of research that relates reading to oral language.

As the present study is only correlational, causal models should be developed to understand the trajectories of lexical and syntactic language development, and the correlates that predict literacy outcomes throughout childhood. Then, perhaps NLP can support just-in-time assessment which can perhaps supplant the high-stakes, multiple-choice assessment practices currently tasked with monitoring children's language and literacy learning. Of course, careful attention must be paid to

the grain-size of such assessment – it is questionable whether the extremely fine grain-size of information produced by the COVFEFE pipeline as described in this study is beneficial for classroom assessment. More research is needed on the combinations of features associated with constructs of interest, and how those can be appropriately reported.

Regardless, my future vision is that just-in-time assessment using these tools can help teachers and students identify which literacy and oral language skills they might focus on to support well-rounded language and literacy learning. It is possible to imagine an educational scenario in the near future in which these technologies enable valid, comprehensive, efficient, and timely diagnostic assessment of students' language and literacy skills at a grain-size appropriate for both teachers and students. If such tools are in classrooms, teachers can further exercise their professional judgment in choosing areas of intervention, and students could be empowered with information about their progress. If valid ML algorithms can be developed to provide feedback to learners and their teachers about spoken and written language development, perhaps a low-cost, effective, low-stakes assessment system could supersede the need for our current high-stakes literacy assessment system. Thus, these technological innovations are more than just cool gadgets: they could actually bring about a democratization of assessment practice that could benefit millions.

References

Abbott, R. D., Berninger, V. W., & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. *Journal of Educational Psychology*, *102*(2), 281.

Adlof, S. M., Catts, H. W., & Little, T. D. (2006). Should the simple view of reading include a fluency component? *Reading and Writing*, *19*(9), 933-958.

Adlof, S. M., Hogan, T. P., & Catts, H. W. (2005). Developmental changes in reading and reading disabilities. In H.W. Catts & A.G. Kamhi (Eds.), *The connections between language and reading disabilities* (pp. 38-51). Mahwah, NJ: Lawrence Erlbaum Associates.

Ahmed, Y., Wagner, R. K., & Lopez, D. (2014). Developmental relations between reading and writing at the word, sentence, and text levels: A latent change score analysis. *Journal of Educational Psychology*, *106*(2), 419.

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson, R. Barr, M. L. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 255–291). New York, NY: Longman.

Asgari, M., Kiss, G., Van Santen, J., Shafran, I., & Song, X. (2014). Automatic measurement of affective valence and arousal in speech. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 965-969). Piscataway, NJ: The Institute of Electrical and Electronics Engineers, Inc. (IEEE).

Babayiğit, S. (2014). The role of oral language skills in reading and listening comprehension of text: A comparison of monolingual (L1) and bilingual (L2) speakers of English language. *Journal of Research in Reading*, *37*(S1), S22-S47.

Babayiğit, S. (2015). The relations between word reading, oral language, and reading comprehension in children who speak English as a first (L1) and second language (L2): A multigroup structural analysis. *Reading and Writing*, *28*(4), 527-544.

Benfatto, M. N., Seimyr, G. Ö., Ygge, J., Pansell, T., Rydberg, A., & Jacobson, C. (2016). Screening for dyslexia using eye tracking during reading. *PloS One, 11*(12), 1-16.

Berninger, V. W., & Abbott, R. D. (2010). Listening comprehension, oral expression, reading comprehension, and written expression: Related yet unique language systems in grades 1, 3, 5, and 7. *Journal of Educational Psychology*, *102*(3), 635-651.

Berninger, V. W., Abbott, R. D., Abbott, S. P., Graham, S., & Richards, T. (2002). Writing and reading: Connections between language by hand and language by eye. *Journal of Learning Disabilities*, *35*(1), 39-56.

Bianco, M., Bressoux, P., Doyen, A. L., Lambert, E., Lima, L., Pellenq, C., & Zorman, M. (2010). Early training in oral comprehension and phonological skills: Results of a three-year longitudinal study. *Scientific Studies of Reading*, *14*(3), 211-246.

Bohn-Gettler, C. M., & Rapp, D. N. (2014). Emotion during reading and writing. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 437-457). New York: Routledge.

Bowey, J. A. (1986). Syntactic awareness in relation to reading skill and ongoing reading comprehension monitoring. *Journal of Experimental Child Psychology, 41*(2), 282-299

Bowey, J. A. (1994). Grammatical awareness and learning to read: A critique. In E. Assink (Ed.), *Literacy acquisition and social context* (pp. 122–149). London: Harvester Wheatsheaf.

Bowyer-Crane, C., Snowling, M. J., Duff, F. J., Fieldsend, E., Carroll, J. M., Miles, J. & Hulme, C. (2008). Improving early language and literacy skills: Differential effects of an oral language versus a phonology with reading intervention. *Journal of Child Psychology and Psychiatry*, *49*(4), 422-432.

Braze, D., Mencl, W. E., Tabor, W., Pugh, K. R., Constable, R. T., Fulbright, R. K., & Shankweiler, D. P. (2011). Unification of sentence processing via ear and eye: An fMRI study. *Cortex*, *47*(4), 416-431.

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5-32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199-231.

Brimo, D., Apel, K., & Fountain, T. (2017). Examining the contributions of syntactic awareness and syntactic knowledge to reading comprehension. *Journal of Research in Reading*, *40*(1), 57-74.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977-990.

Buchweitz, A., Mason, R. A., Tomitch, L. M., & Just, M. A. (2009). Brain activation for reading and listening comprehension: An fMRI study of modality effects and individual differences in language comprehension. *Psychology & Neuroscience*, *2*(2), 111–123.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine*, *25*(3), 27-27.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay scoring: Current applications and future directions* (pp. 55–67). New York: Routledge.

Cain, K. (2007). Syntactic awareness and reading ability: Is there any evidence for a special relationship? *Applied Psycholinguistics*, *28*(4), 679-694.

Cain, K. (2015). Literacy development: The interdependent roles of oral language and reading comprehension. In R. H. Bahr, & E. R. Silliman (Eds.), *Routledge handbook of communication disorders* (pp. 204-214). New York, NY: Routledge

Cain, K., & Oakhill, J. (2006). Profiles of children with specific reading comprehension difficulties. *British Journal of Educational Psychology*, *76*(4), 683-696.

Cain, K., & Oakhill, J. (2014). Reading comprehension and vocabulary: Is vocabulary more important for some aspects of comprehension? *L'Annee Psychologique*, *114*(4), 647-662.

Cameron, C. A., Lee, K., Webster, S., Munro, K., Hunt, A. K., & Linton, M. J. (1995). Text cohesion in children's narrative writing. *Applied Psycholinguistics*, *16*(3), 257-269.

Carretti, B., Motta, E., & Re, A. M. (2016). Oral and written expression in children with reading comprehension difficulties. *Journal of Learning Disabilities*, *49*(1), 65-76.

Casas, B., Català, N., Ferrer-i-Cancho, R., Hernández-Fernández, A., & Baixeries, J. (2018). The polysemy of the words that children learn over time. *Interaction Studies*, *19*(3), 389-426.

Catts, H. W., Fey, M. E., Tomblin, J. B., & Zhang, X. (2002). A longitudinal investigation of reading outcomes in children with language impairments. *Journal of Speech, Language, and Hearing Research, 45*(6), 1142-1157.

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (1999). Language basis of reading and reading disabilities: Evidence from a longitudinal investigation. *Scientific Studies of Reading*, *3*(4), 331-361.

Catts, H. W., Hogan, T. P., & Adolf, S. M. (2005). Developmental changes in reading and reading disabilities. In H. W. Catts & A.G. Kamhi, (Eds.) *The connections between language and reading disabilities* (pp. 23-36). Mahwah, NJ: Erlbaum.

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement, 76*(1), 114-140.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences, 10*(7), 335-344.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences, 10*(7), 287-291.

Chen, R. S., & Vellutino, F. R. (1997). Prediction of reading ability: A cross-validation study of the simple view of reading. *Journal of Literacy Research, 29*(1), 1-24.

Chinaei, H., Currie, L. C., Danks, A., Lin, H., Mehta, T., & Rudzicz, F. (2017). Identifying and avoiding confusion in dialogue with people with Alzheimer's disease. *Computational* Linguistics, 43(2), 377-406.

Chipere, N., Malvern, D., Richards, B., & Duran, P. (2001). Using a corpus of school children's writing to investigate the development of vocabulary diversity. In P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja, (Eds). *Technical Papers. Volume 13. Special Issue: Proceedings of the Corpus Linguistics 2001 Conference* (pp. 126-133). Lancaster University, University Centre for Computer Corpus Research on Language.

Clarke, P. J., Snowling, M. J., Truelove, E., & Hulme, C. (2010). Ameliorating children's reading-comprehension difficulties: A randomized controlled trial. *Psychological Science*, *21*(8), 1106-1116.

Coker, Jr., D. L., Jennings, A. S., Farley-Ripple, E., & MacArthur, C. A. (2018). The type of writing instruction and practice matters: The direct and indirect effects of writing instruction and student practice on reading achievement. *Journal of Educational Psychology*, *110*(4), 502.

Coltheart, V., Laxon, V. J., & Keating, C. (1988). Effects of word imageability and age of acquisition on children's reading. *British Journal of Psychology, 79*(1), 1-12.

Cope, B., & Kalantzis, M. (Eds.). (2000). *Multiliteracies: Literacy learning and the design of social futures*. New York: Routledge.

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.

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, *26*, 66-79.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing, 32*, 1-16.

Crossley, S. A., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, *60*(3), 573-605.

Crossley, S. A., Weston, J. L., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28*(3), 282-311.

Davidson, D., Raschke, V. R., & Pervez, J. (2010). Syntactic awareness in young monolingual and bilingual (Urdu–English) children. *Cognitive Development*, *25*(2), 166-182.

Deacon, S. H., & Kieffer, M. (2018). Understanding how syntactic awareness contributes to reading comprehension: Evidence from mediation and longitudinal models. *Journal of Educational Psychology*, *110*(1), 72.

Deborah, L. J., Baskaran, R., & Kannan, A. (2010). A survey on internal validity measure for cluster validation. *International Journal of Computer Science & Engineering Survey*, *1*(2), 85-102.

Decker, S. L., Roberts, A. M., Roberts, K. L., Stafford, A. L., & Eckert, M. A. (2016). Cognitive components of developmental writing skill. *Psychology in the Schools*, *53*(6), 617-625.

Demont, E., & Gombert, J. E. (1996). Phonological awareness as a predictor of recoding skills and syntactic awareness as a predictor of comprehension skills. *British Journal of Educational Psychology*, *66*(3), 315-332.

Dickinson, D. K., Golinkoff, R. M., & Hirsh-Pasek, K. (2010). Speaking out for language: Why language is central to reading development. *Educational Researcher*, *39*(4), 305-310.

Dockrell, J. E., & Connelly, V. (2015). The role of oral language in underpinning the text generation difficulties in children with specific language impairment. *Journal of Research in Reading*, *38*(1), 18-34.

Dunst, C. J., Trivette, C. M., Masiello, T., Roper, N., & Robyak, A. (2006). Framework for developing evidence-based early literacy learning practices. *CELL (Centre for Early Literacy Learning) Papers*, *1*(1), 1-12.

Durrant, P., & Brenchley, M. (2018). Development of vocabulary sophistication across genres in English children's writing. *Reading and Writing*, 1-27. DOI 10.1007/s11145-018-9932-8.

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness*, *2*(1), 1-44.

Evanini, K., Heilman, M., Wang, X., & Blanchard, D. (2015). Automated scoring for the TOEFL Junior® comprehensive writing and speaking test. *ETS Research Report Series*, *2015*(1), 1-11.

Fang, Z., Schleppegrell, M. J., & Cox, B. E. (2006). Understanding the language demands of schooling: Nouns in academic registers. *Journal of Literacy Research, 38*(3), 247-273.

Fitzgerald, J., & Shanahan, T. (2000). Reading and writing relations and their development. *Educational Psychologist*, *35*(1), 39-50.

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2018). *Learning disabilities: From identification to intervention* (2nd edition). New York: The Guilford Press.

Foorman, B. R., Koon, S., Petscher, Y., Mitchell, A., & Truckenmiller, A. (2015). Examining general and specific factors in the dimensionality of oral language and reading in 4th–10th grades. *Journal of Educational Psychology*, *107*(3), 884.

Fraser, K. (2016). *Automatic text and speech processing for the detection of dementia* (Doctoral dissertation, University of Toronto.

Fraser, K. C., Rudzicz, F., & Rochon, E. (2013). Using text and acoustic features to diagnose progressive aphasia and its subtypes. In F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, & P. Perrier (Eds.), *INTERSPEECH* (pp. 2177-2181). ISCA Archive, http://www.isca-speech.org/archive/interspeech_2013.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education*, *9*(2), 20-28.

Galloway, E. P., & Uccelli, P. (2019). Beyond reading comprehension: exploring the additional contribution of Core Academic Language Skills to early adolescents' written summaries. *Reading and Writing*, *32*(3), 729-759.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczaj (Ed.), *Language development: Vol. 2. Language, thought, and culture* (pp. 301-334). Hillsdale, NJ: Erlbaum.

Gentner, D. (2006). Why verbs are hard to learn. In K. Hirsh-Pasek, & R. Golinkoff, (Eds.) *Action meets word: How children learn verbs* (pp. 544-564). Oxford University Press

Gernsbacher, M. A., Varner, K. R., & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 430.

Geva, E., & Herbert, K. (2013). Assessment and interventions for English language learners with learning disabilities. In B. Wong & D. Butler, (Eds.), *Learning about Learning Disabilities* (4th Edition, pp. 271-198). Amsterdam: Elsevier.

Geva, E., & Farnia, F. (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Reading and Writing, 25*(8), 1819-1845.

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation, 12*(4), 395-427.

Gottardo, A. (2002). The relationship between language and reading skills in bilingual Spanish-English speakers. *Topics in Language Disorders*, *22*(5), 46-70.

Gottardo, A., Mirza, A., Koh, P. W., Ferreira, A., & Javier, C. (2018). Unpacking listening comprehension: the role of vocabulary, morphological awareness, and syntactic knowledge in reading comprehension. *Reading and Writing*, *31*(8), 1741-1764.

Gottardo, A., Stanovich, K. E., & Siegel, L. S. (1996). The relationships between phonological sensitivity, syntactic processing, and verbal working memory in the reading performance of third-grade children. *Journal of Experimental Child Psychology*, *63*(3), 563-582.

Gough, P. B., Hoover, W., & Peterson, C. L. (1996). Some observations on the simple view of reading. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties* (pp. 1–13). Mahwah, NJ: Lawrence Erlbaum Associates, Inc

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*(1), 6-10.

Graham, S., & Hebert, M. (2011). Writing to read: A meta-analysis of the impact of writing and writing instruction on reading. *Harvard Educational Review*, *81*(4), 710-744.

Graham, S., Liu, X., Aitken, A., Ng, C., Bartlett, B., Harris, K. R., & Holzapfel, J. (2018). Effectiveness of literacy programs balancing reading and writing instruction: A meta-analysis. *Reading Research Quarterly*, *53*(3), 279-304.

Gray, S., Catts, H., Logan, J., & Pentimonti, J. (2017). Oral language and listening comprehension: Same or different constructs? *Journal of Speech, Language, and Hearing Research*, *60*(5), 1273-1284.

Haft, S. L., Duong, P. H., Ho, T. C., Hendren, R. L., & Hoeft, F. (2019). Anxiety and attentional bias in children with specific learning disorders. *Journal of Abnormal Child Psychology, 47*(3), 487-497.

Hagtvet, B. E. (2003). Listening comprehension and reading comprehension in poor decoders: Evidence for the importance of syntactic and semantic skills as well as phonological skills. *Reading and Writing*, *16*(6), 505-539.

Halliday, M. A. (1993). Towards a language-based theory of learning. *Linguistics and Education, 5*(2), 93-116.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H Brookes Publishing.

Harris, S. (2018). Identifying student difficulty and frustration from discussion forum postings. Unpublished Master of Science thesis. Athabasca University.

Hassanali, K. N., Yoon, S. Y., & Chen, L. (2015). Automatic scoring of non-native children's spoken language proficiency. In S. Steidl, A. Bat–iner, & O. Jokisch (Eds.), *SLaTE 2015 - Workshop on Speech and Language Technology in Education* (pp. 13-18). International Speech Communication Association Archive, http://www.isca-speech.org/archive/slate_2015.

Hodges, T. S., Feng, L., Kuo, L. J., & McTigue, E. (2016). Discovering the literacy gap: A systematic review of reading and writing theories in research. *Cogent Education*, *3*(1), 1-13.

Hoffman, L. M., Loeb, D. F., Brandel, J., & Gillam, R. B. (2011). Concurrent and construct validity of oral language measures with school-age children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *54*, 1597–1608.

Hogan, T. P., Adlof, S. M., & Alonzo, C. N. (2014). On the importance of listening comprehension. *International Journal of Speech-Language Pathology*, *16*(3), 199-207.

Holahan, J. M., Ferrer, E., Shaywitz, B. A., Rock, D. A., Kirsch, I. S., Yamamoto, K., & Shaywitz, S. E. (2018). Growth in reading comprehension and verbal ability from grades 1 through 9. *Journal of Psychoeducational Assessment*, *36*(4), 307-321.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127-160.

Hsu, C. K., Hwang, G. J., & Chang, C. K. (2010). Development of a reading material recommendation system based on a knowledge engineering approach. *Computers & Education, 55*(1), 76-83.

Humphry, S., Heldsinger, S., & Dawkins, S. (2017). A two-stage assessment method for assessing oral language in early childhood. *Australian Journal of Education*, *61*(2), 124-140.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: Springer.

Jang, E. E. (2019). Potential of machine learning approaches for advancing diagnostic language a[ss]essment. Keynote presentation at the 9th Canadian Association of Language Assessment (CALA) Symposium. Toronto, ON, Canada.

Jang, E. E., Cummins, J., Wagner, M., Stille, S., & Dunlop, M. (2015). Investigating the homogeneity and distinguishability of STEP proficiency descriptors in assessing English language learners in Ontario schools. *Language Assessment Quarterly, 12*(1), 87-109.

Jiang, N. (2000). Lexical representation and development in a second language. Applied linguistics, 21(1), 47-77.

Kapantzoglou, M., Fergadiotis, G., & Auza Buenavides, A. (2019). Psychometric evaluation of lexical diversity indices in Spanish narrative samples from children with and without developmental language disorder. *Journal of Speech, Language, and Hearing Research, 62*(1), 70-83.

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, *47*(7), 451-464.

Kendeou, P., Bohn-Gettler, C., White, M., & Van Den Broek, P. (2008). Children's inference generation across different media. *Journal of Research in Reading*, *31*(3), 259-272.

Kendeou, P., Van den Broek, P., White, M. J., & Lynch, J. S. (2009). Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills. *Journal of Educational Psychology*, *101*(4), 765.

Kim, M., Crossley, S. A., & Skalicky, S. (2018). Effects of lexical features, textual properties, and individual differences on word processing times during second language reading comprehension. *Reading and Writing*, *31*(5), 1155–1180

Kim, Y. H., & Jang, E. E. (2009). Differential functioning of reading subskills on the OSSLT for L1 and ELL students: A multidimensionality model-based DBF/DIF approach. *Language Learning*, *59*(4), 825-865.

Kim, Y. S. G. (2017). Why the simple view of reading is not simplistic: Unpacking component skills of reading using a direct and indirect effect model of reading (DIER). *Scientific Studies of Reading*, *21*(4), 310-333.

Kim, Y. S. G., Park, C., & Park, Y. (2015). Dimensions of discourse level oral language skills and their relation to reading comprehension and written composition: An exploratory study. *Reading and Writing*, *28*(5), 633-654.

Kim, Y. S., Puranik, C., & Otaiba, S. A. (2015). Developmental trajectories of writing skills in first grade: Examining the effects of SES and language and/or speech impairments. *The Elementary School Journal*, *115*(4), 593-613.

Kincaid, J., Fishburne, R., Rogers, R., & Chissom, B. (1975). Derivation of new readability formulas (Automated Readability Index, fog count, and Flesch Reading Ease Formula) for Navy enlisted personnel. Navy Training Command Research Branch Report 8-75

Kintsch, E., & Kintsch, W. (2005). Comprehension. In S. Paris & S. A. Stahl, (Eds.), *Children's reading comprehension and assessment* (pp. 89-110). Mahwah, NJ: Lawrence Erlbaum.

Kintsch, W. (1992). A cognitive architecture for comprehension. In H. L. Pick, Jr., P. W. van den Broek, & D. C. Knill (Eds.), *Cognition: Conceptual and methodological issues* (pp. 143-163). Washington, DC: American Psychological Association.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, *49*(4), 294.

Kintsch, W. (2012). Psychological models of reading comprehension and their implications for assessment. In J. Sabatini, E. Albro, & T. O'Reilly, (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 21-38). Lanham, MD: Rowman & Littlefield Education.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363-394.

Kirby, J. R., & Savage, R. S. (2008). Can the simple view deal with the complexities of reading? *Literacy*, *42*(2), 75-82.

Klingner, J. K. (2004). Assessing reading comprehension. *Assessment for Effective Intervention*, *29*(4), 59-70.

Komeili, M., Pou-Prom, C., Liaqat, D., Fraser, K. C., Yancheva, M., & Rudzicz, F. (2019). Talk2Me: Automated linguistic data collection for personal assessment. *PloS one*, *14*(3), e0212342.

Koo, J., Becker, B. J., & Kim, Y. S. (2014). Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach. *Language Testing, 31*(1), 89–109

Korobov, M. & Lopuhin, K. (2019). ELI5 Documentation (Release 0.8.2).

Krishnan, S., Watkins, K. E., & Bishop, D. V. (2016). Neurobiological basis of language learning difficulties. *Trends in Cognitive Sciences*, *20*(9), 701-714.

Law, J., Rush, R., King, T., Westrupp, E., & Reilly, S. (2018). Early home activities and oral language skills in middle childhood: A quantile analysis. *Child Development*, *89*(1), 295-309.

Lee, P. Hu, Y. Chen, K., Tarn, J. M., & Cheng, L. (2018). Cyberbullying detection on social network services. *Proceedings of Twenty-Second Pacific Asia Conference on Information Systems*. Association for Information Systems. Accessed on August 1, 2019 from https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1060&context=pacis2018,

Lesaux, N. K., Crosson, A. C., Kieffer, M. J., & Pierce, M. (2010). Uneven profiles: Language minority learners' word reading, vocabulary, and reading comprehension skills. *Journal of Applied Developmental Psychology*, *31*(6), 475-483.

Lesaux, N. K., Geva, E., Koda, K., Siegel, L. S., & Shanahan, T. (2008). Development of literacy in second language learners. In D. August & T. Shanahan (Eds.), *Developing reading and writing in second language learners*, (pp. 27–59). New York: Center for Applied Linguistics and International Reading Association.

Lesaux, N. K., & Harris, J. R. (2013). Linguistically diverse students' reading difficulties: Implications for models of learning disabilities identification and effective instruction. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 69-84). New York: Guilford Press.

Lesaux, N. K., & Kieffer, M. J. (2010). Exploring sources of reading comprehension difficulties among language minority learners and their classmates in early adolescence. *American Educational Research Journal*, *47*(3), 596-632.

Lesaux, N. K., Lipka, O., & Siegel, L. S. (2006). Investigating cognitive and linguistic abilities that influence the reading comprehension skills of children from diverse linguistic backgrounds. *Reading and Writing*, *19*(1), 99-131.

Lesaux, N. K., Rupp, A. A., & Siegel, L. S. (2007). Growth in reading skills of children from diverse linguistic backgrounds: Findings from a 5-year longitudinal study. *Journal of Educational Psychology*, *99*(4), 821-834.

Lewis, W. D. (2002). Measuring conceptual distance using WordNet: The design of a metric for measuring semantic similarity. *Coyote Papers: Working Papers in Linguistics, Language in Cognitive Science*, *12*, 10-16.

Lo, Y. C., Rensi, S. E., Torng, W., & Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, *23*(8), 1538-1546.

Meng, L., Huang, R., & Gu, J. (2013). A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology*, *6*(1), 1-12.

Lonigan, C. J., & Milburn, T. F. (2017). Identifying the dimensionality of oral language skills of children with typical development in preschool through fifth grade. *Journal of Speech, Language, and Hearing Research*, *60*(8), 2185-2198.

Lu, X. (2009). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics, 14*(1), 3-28.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474-496.

Madnani, N., Loukina, A., von Davier, A., Burstein, J., & Cahill, A. (2017). Building better open-source tools to support fairness in automated scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 41-52). Stroudsburg, PA: Association for Computational Linguistics.

Manning, C. D. (2003). Probabilistic syntax. In Bod, R., Hay, J., & Jannedy, S. (Eds.), *Probabilistic linguistics* (pp. 289-342). Cambridge, MA: MIT Press.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., & McClosky, D. (2014). The Stanford coreNLP natural langua[ge] processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (*pp. 55–60). Stroudsburg, PA: Association for Computational Linguistics.

McNamara, D. S., & Allen, L. K. (2017). Toward an integrated perspective of writing as a discourse process. In M. F. Schober, D. N. Rapp, & M. A. Britt (Eds.), *Routledge Handbook of Discourse Processes* (pp. 362-389). New York: Routledge.

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*(2), 499-515.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6-20.

Napolitano, D., Sheehan, K., & Mundkowsky, R. (2015, June). Online readability and text complexity analysis with TextEvaluator. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (pp. 96-100).

Nation, K., Clarke, P., Marshall, C. M., & Durand, M. (2004). Hidden language impairments in children. *Journal of Speech, Language, and Hearing Research*, *47*, 199-211.

Nation, K., & Snowling, M. J. (2000). Factors influencing syntactic awareness skills in normal readers and poor comprehenders. *Applied Psycholinguistics*, *21*(2), 229-241.

Nation, K., & Snowling, M. J. (2004). Beyond phonological skills: Broader language skills contribute to the development of reading. *Journal of Research in Reading*, *27*(4), 342-356.

NICHD Early Child Care Research Network. (2005). Pathways to reading: The role of oral language in the transition to reading. *Developmental Psychology*, *41*(2), 428-442.

Oakhill, J., Cain, K., & Yuill, N. (1998). Individual differences in children's comprehension skill: Toward an integrated model. In C. Hulme & R. Malatesha Joshi (Eds.), *Reading and spelling: Development and disorders* (pp. 343-367). New York: Routledge.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and Demonstration of Multidimensional IRT-Based Internal Measures of Differential Functioning of Items and Tests. *Journal of Educational Measurement*, *34*(3), 253-272.

Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology*, *98*(3), 554.

Ouellette, G., & Beers, A. (2010). A not-so-simple view of reading: How oral vocabulary and visual-word recognition complicate the story. *Reading and Writing*, *23*(2), 189-208.

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices—past, present, and future. In S. G. Paris & S. A. Stahl, (Eds.), *Children's reading comprehension and assessment* (pp. 31-88). Mahwah, NJ: Lawrence Erlbaum.

Pearson, P. D., Valencia, S. W., & Wixson, K. (2014). Complicating the world of reading assessment: Toward better assessments for better teaching. *Theory into Practice*, *53*(3), 236-246.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004* (pp. 38-41). Stroudsburg, PA: Association for Computational Linguistics.

Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357-383.

Perfetti, C. A., & Adlof, S. M. (2012). Reading comprehension: A conceptual framework from word meaning to text meaning. In J. Sabatini, E. Albro, & T. O'Reilly, (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 3-20). Lanham, MD: Rowman & Littlefield Education.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme, (Eds.), *The science of reading: A handbook* (pp. 227-247). Malden, MA: Blackwell.

Ponari, M., Norbury, C. F., & Vigliocco, G. (2018). Acquisition of abstract concepts is influenced by emotional valence. *Developmental Science, 21*(2), e12549.

Poulsen, M., & Gravgaard, A. K. (2016). Who did what to whom? The relationship between syntactic aspects of sentence comprehension and text comprehension. *Scientific Studies of Reading*, *20*(4), 325-338.

Poznyk, D. (2018). Mplus. In B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. Los Angeles, CA: Sage Reference.

Prevoo, M. J., Malda, M., Mesman, J., & van IJzendoorn, M. H. (2016). Within-and cross-language relations between oral language proficiency and school outcomes in bilingual children with an immigrant background: A meta-analytical study. *Review of Educational Research*, *86*(1), 237-276.

Proctor, C. P., Silverman, R. D., Harring, J. R., & Montecillo, C. (2012). The role of vocabulary depth in predicting reading comprehension among English monolingual and Spanish–English bilingual children in elementary school. *Reading and Writing*, *25*(7), 1635-1664.

Puranik, C. S., & Lonigan, C. J. (2012). Early writing deficits in preschoolers with oral language difficulties. *Journal of Learning Disabilities*, *45*(2), 179-190.

Quinn, J. M., Wagner, R. K., Petscher, Y., & Lopez, D. (2015). Developmental relations between vocabulary knowledge and reading comprehension: A latent change score modeling study. *Child Development*, *86*(1), 159-175.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Rich, E. (1985). Artificial intelligence and the humanities. *Computers and the Humanities, 19*(2), 117-122.

Ricketts, J., Nation, K., & Bishop, D. V. (2007). Vocabulary is important for some, but not all reading skills. *Scientific Studies of Reading*, *11*(3), 235-257.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, *23*(4), 441-474.

Saez, Y., Baldominos, A., & Isasi, P. (2017). A comparison study of classifier algorithms for cross-person physical activity recognition. *Sensors*, 17(1), 66.

Scarborough, H. (2005). Developmental relationships between language and reading: Reconciling a beautiful hypothesis with some ugly facts. In Catts, H. W., & Kamhi, A. G. (Eds.), *The connections between language and reading disabilities*. Mahwah, NJ: Lawrence Erlbaum.

Shanahan, T. (2016). Relationships between reading and writing development. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd Ed., pp. 194-207). New York: Guilford.

Shapiro, E. S., Fritschmann, N. S., Thomas, L. B., Hughes, C. L., & McDougal, J. (2014). Concurrent and predictive validity of reading retell as a brief measure of reading comprehension for narrative text. *Reading Psychology*, *35*(7), 644-665.

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, *24*(1), 99-128.

Siegel, L. S. (2008). Morphological awareness skills of English language learners and children with dyslexia. *Topics in Language Disorders*, *28*(1), 15-27.

Silverman, R. D., Speece, D. L., Harring, J. R., & Ritchey, K. D. (2013). Fluency has a role in the simple view of reading. *Scientific Studies of Reading*, *17*(2), 108-133.

Simonnet, D., Girard, N., Anquetil, E., Renault, M., & Thomas, S. (2019). Evaluation of children cursive handwritten words for e-education. *Pattern Recognition Letters*, *121*, 133-139.

Snefjella, B., & Kuperman, V. (2016). It's all in the delivery: Effects of context valence, arousal, and concreteness on visual word processing. *Cognition, 156*, 135-146.

Snow, C. E., & Biancarosa, G. (2003). *Adolescent literacy and the achievement gap: What do we know and where do we go from here?* New York, NY: Carnegie Corporation.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1631-1642). Stroudsburg, PA: Association for Computational Linguistics.

Somasundaran, S., Lee, C. M., Chodorow, M., & Wang, X. (2015). Automated scoring of picture-based story narration. In *Proceedings of the Tenth Workshop on Innovative Use of*

NLP for Building Educational Applications (pp. 42-48). Stroudsburg, PA: Association for Computational Linguistics.

Spencer, S., Clegg, J., Stackhouse, J., & Rush, R. (2017). Contribution of spoken language and socio-economic background to adolescents' educational achievement at age 16 years. *International Journal of Language & Communication Disorders*, *52*(2), 184-196.

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, *38*(4), 598-605.

Stage, S. A., Abbott, R. D., Jenkins, J. R., & Berninger, V. W. (2003). Predicting response to early reading intervention from verbal IQ reading-related language abilities, attention ratings, and verbal IQ-word reading discrepancy: Failure to validate discrepancy method. *Journal of Learning Disabilities, 36*, 24-33.

Stanovich, K. E. (2005). The future of a mistake: Will discrepancy measurement continue to make the learning disabilities field a pseudoscience? *Learning Disability Quarterly*, *28*(2), 103-106.

Stanovich, K. E., & Siegel, L. S. (1994). Phenotypic performance profile of children with reading disabilities: A regression-based test of the phonological-core variable-difference model. *Journal of Educational Psychology*, *86*(1), 24.

Stein, C. L., Cairns, H. S., & Zurif, E. B. (1984). Sentence comprehension limitations related to syntactic deficits in reading-disabled children. *Applied Psycholinguistics*, *5*(4), 305-322.

Suggate, S. P. (2016). A meta-analysis of the long-term effects of phonemic awareness, phonics, fluency, and reading comprehension interventions. *Journal of Learning Disabilities*, *49*(1), 77-96.

Tannenbaum, K. R., Torgesen, J. K., & Wagner, R. K. (2006). Relationships between word knowledge and reading comprehension in third-grade children. *Scientific Studies of Reading*, *10*(4), 381-398.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

Tilstra, J., McMaster, K., Van den Broek, P., Kendeou, P., & Rapp, D. (2009). Simple but complex: Components of the simple view of reading across grade levels. *Journal of Research in Reading*, *32*(4), 383-401.

Tomblin, J. B., & Zhang, X. (2006). The dimensionality of language ability in school-age children. *Journal of Speech, Language, and Hearing Research, 49*, 1193-1208.

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. In *2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018),* arXiv:1807.01308.

Tong, X., Deacon, S. H., & Cain, K. (2014). Morphological and syntactic awareness in poor comprehenders: Another piece of the puzzle. Journal of Learning Disabilities, 47(1), 22-33.

Tong, X., & McBride, C. (2016). Reading comprehension mediates the relationship between syntactic awareness and writing composition in children: A longitudinal study. *Journal of Psycholinguistic Research*, *45*(6), 1265-1285.

Tosto, M. G., Hayiou-Thomas, M. E., Harlaar, N., Prom-Wormley, E., Dale, P. S., & Plomin, R. (2017). The genetic architecture of oral language, reading fluency, and reading comprehension: A twin study from 7 to 16 years. *Developmental Psychology*, *53*(6), 1115-1129.

Tunmer, W. E., & Chapman, J. W. (2012). The simple view of reading redux: Vocabulary knowledge and the independent components hypothesis. *Journal of Learning Disabilities*, *45*(5), 453-466.

Verhoeven, L., van Leeuwe, J., & Vermeer, A. (2011). Vocabulary growth and reading development across the elementary school years. *Scientific Studies of Reading*, *15*(1), 8-25.

Walczak, N., Fasching, J., Cullen, K., Morellas, V., & Papanikolopoulos, N. (2018). Toward identifying behavioral risk markers for mental health disorders: an assistive system for monitoring children's movements in a preschool classroom. *Machine Vision and Applications*, *29*(4), 703-717.

Walford, G., Tucker, E., & Viswanathan, M. (2010). *The SAGE handbook of measurement*. Sage Publications.

Walley, A. C., Metsala, J. L., & Garlock, V. M. (2003). Spoken vocabulary growth: Its role in the development of phoneme awareness and early reading ability. *Reading and Writing*, *16*(1-2), 5-20.

Wang, T., & Hirst, G. (2011). Refining the notions of depth and density in WordNet-based semantic similarity measures. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1003-1011). Stroudsburg, PA: Association for Computational Linguistics.

Warriner, A. B., & Kuperman, V. (2015). Affective biases in English are bi-dimensional. *Cognition and Emotion, 29*(7), 1147-1167.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191-1207.

Wiebe, J., & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 486-497). Berlin: Springer.

Wilson T., Wiebe J., & Hoffmann P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP),* (pp. 347–354). Stroudsburg, PA: Association for Computational Linguistics.

Wolf, M. C., Muijselaar, M. M. L., Boonstra, A. M., & De Bree, E. H. (2018). The relationship between reading and listening comprehension: shared and modality-specific components. *Reading and Writing*, 1-21. DOI 10.1007/s11145-018-9924-8

Wood, C. L., Bustamante, K. N., Schatschneider, C., & Hart, S. (2018). Relationship between children's lexical diversity in written narratives and performance on a standardized reading vocabulary measure. *Assessment for Effective Intervention*, *44*(3), 173-183.

Wright, M. N., Wager, S., & Probst, P. (2019). Package 'ranger'.

Wright, M. N., Ziegler, A., & König, I. R. (2016). Do little interactions get lost in dark random forests?. *BMC Bioinformatics*, *17*(1), 145.

Yancheva, M., & Rudzicz, F. (2013). Automatic detection of deception in child-produced speech using syntactic complexity features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 944-953). Stroudsburg, PA: Association for Computational Linguistics.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society, 104*(5), 444-466.

Zechner, K., Evanini, K., Yoon, S. Y., Davis, L., Wang, X., Chen, L., & Leong, C. W. (2014). Automated scoring of speaking items in an assessment for teachers of English as a Foreign Language. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 134-142). Stroudsburg, PA: Association for Computational Linguistics.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, *51*(10), 883-895.

Zhang, M., Dorans, N., Li, C., & Rupp, A. (2017). Differential feature functioning in automated essay scoring. In H. Jiao & R. W. Lissitz, (Eds.), *Test fairness in the new generation of large-scale assessment,* (pp. 185-208). Charlotte, NC: Information Age.

Appendix A

Table 34

Pairwise correlations between the outcome variable (reading comprehension score) and each linguistic feature extracted through NLP

Linguistic feature extracted by NLP	Oral elicitation, regular dataset	Oral elicitation, modified dataset	Text elicitation, regular dataset	Text elicitation, modified dataset
Word-level grammatical constituents				
Number of words	.09	.17	.34	-.06
Adjectives	**.01**	**.35**	**.01**	**.09**
Adverbs	.32	-.02	-.06	-.13
Coordinates	.06	-.17	.04	-.18
Demonstratives	.03	.03	.28	-.10
Determiners	-.09	-.17	.13	.15
Inflected verbs	.20	.04	<.01	.18
Light verbs (be, have, come, go, give, take, make, do, get, move, put)	-.01	<.01	.04	.02
Nouns	-.14	-.04	-.03	.05
Function words[a]	.10	-.06	.06	-.07
Prepositions	**.07**	**.12**	**.09**	**.03**
Personal pronouns	.18	.06	-.14	-.13
Subordinating conjunctions	**.23**	**.10**	**.18**	**.13**
Verbs	.25	.08	-.16	.06
Ratio of nouns to nouns and verbs	-.24	-.07	.12	<.01
Ratio of nouns to verbs	-.15	.14	.05	-.06
Ratio of Personal pronouns to Personal pronouns and nouns	.16	.07	-.09	-.11
Ratio of subordinating to coordinating conjunctions	**.10**	**.02**	**.10**	**.18**
Propositional density: verbs adjectives adverbs prepositions " + conjunctions)/numWords	.31	.15	-.07	-.04
Phrase-level grammatical constituents				
Verb phrases	.24	.05	-.44	-.05
Adjectival phrase consisting of an adjective	-.01	.02	-.05	.13
Adjectival phrase consisting of an adjective and prepositional phrase	.08	.1	.03	-.19
Adjectival phrase consisting of an adverb and adjectival phrase	.06	-.03	.02	<.01
Adverbial phrase consisting of an adverb	.22	.09	.01	-.02

Adverbial phrase consisting of an adverb and prepositional phrase	.16	.01	-.06	n/a
Adverbial phrase consisting of two adverbs	.03	.12	.11	-.16
Interjection phrase consisting of an interjection	-.16	-.01	n/a	n/a
Noun phrase consisting of a cardinal number	.12	-.12	-.02	-.17
Noun phrase consisting of a cardinal number and a singular or mass noun	-.20	-.10	.18	.02
Noun phrase consisting of a cardinal number and a plural noun	-.10	-.07	-.06	.02
Noun phrase consisting of a determiner	.06	.09	.16	-.14
Noun phrase consisting of a determiner, another determiner, and a noun	n/a	.16	n/a	n/a
Noun phrase consisting of a determiner, adjective, and a singular or mass noun	.09	.03	-.06	.23
Noun phrase consisting of a determiner, adjective, and a plural noun	n/a	n/a	.24	.08
Noun phrase consisting of a determiner and a singular or mass noun	-.13	-.22	.09	.06
Noun phrase consisting of a determiner and two singular or mass nouns	-.06	-.02	.06	.13
Noun phrase consisting of a determiner and a plural noun	.18	-.21	.05	.04
Noun phrase consisting of an existential “there”	-.03	-.04	-.08	-.12
Noun phrase consisting of a foreign word	-.03	.12	-.04	.08
Noun phrase consisting of an adjective and a singular or mass noun	<.01	.07	-.23	-.02
Noun phrase consisting of an adjective and a plural noun	-.13	.01	.21	-.14
Noun phrase consisting of a singular or mass noun	.10	.12	-.14	.19
Noun phrase consisting of two singular or mass nouns	.10	.35	-.04	-.12
Noun phrase consisting of a singular or mass noun and a plural noun	-.01	-.10	.04	.01
Noun phrase consisting of a proper singular noun	.09	.08	.03	<.01
Noun phrase consisting of a plural noun	.06	.10	-.04	-.09
Noun phrase consisting of a noun phrase, coordinating conjunction, and noun phrase	**-.32**	**-.11**	**-.21**	**-.14**
Noun phrase consisting of two noun phrases	.01	-.11	-.09	-.07
Noun phrase consisting of a noun phrase and prepositional phrase	-.09	-.01	.26	.08
Noun phrase consisting of a noun phrase and a subordinated clause	<.01	-.02	.02	.02
Noun phrase consisting of a noun phrase and verb phrase	-.27	-.05	.15	.20
Noun phrase consisting of a personal pronoun	.21	-.01	-.19	-.16
Noun phrase consisting of possessive pronoun, adjective, and noun	-.14	.17	.11	.16

Noun phrase consisting of possessive pronoun and noun	-.08	.04	.02	.03
Noun phrase consisting of possessive pronoun and plural noun	.07	-.03	-.03	-.24
Noun phrase consisting of an adverb	.02	.26	-.15	.16
Particle phrase consisting of a particle	-.03	-.01	.03	.20
Prepositional phrase consisting of a preposition or subordinating conjunction	-.24	-.07	-.08	.14
Prepositional phrase consisting of a preposition and noun phrase	**.04**	**.15**	**.23**	**.10**
Prepositional phrase consisting of a preposition or subordinating conjunction and a prepositional phrase	.21	<.01	n/a	n/a
Prepositional phrase consisting of a preposition and a simple declarative clause	.04	-.01	-.04	.02
Prepositional phrase consisting of "to" and a noun phrase	-.04	.16	-.06	.08
Root consisting of a fragment	.04	n/a	-.10	.09
Root consisting of a noun phrase	n/a	n/a	-.26	.20
Root consisting of a simple declarative clause	-.06	.01	-.23	-.26
Root consisting of a direct question introduced by a *wh*-word or a *wh*-phrase.	n/a	n/a	.12	.22
Simple declarative clause	.04	-.11	.34	-.11
Simple declarative clause consisting of an adverb, noun phrase, and verb phrase	n/a	n/a	.03	.06
Subordinating conjunction followed by a preposition and simple declarative clause	.17	.03	-.02	-.23
Subordinating conjunction followed by a simple declarative clause	-.05	-.01	-.14	.16
Subordinating conjunction followed by a Wh-adverbial phrase and a simple declarative clause	.13	-.05	.02	-.02
Subordinating conjunction followed by a Wh-noun phrase and a simple declarative clause	.25	.08	.22	-.02
Simple declarative clause consisting of coordinating conjunction, noun phrase, and verb phrase	-.17	-.03	.04	.05
Wh-adverb Phrase consisting of a Wh-adverb	.15	-.07	.12	.14
Wh-noun phrase consisting of a Wh-determiner	.19	.01	.25	-.08
Wh-noun phrase consisting of a wh-pronoun	**.21**	**.02**	**.07**	**.03**
Verb phrase consisting of a modal and verb phrase	-.06	-.07	.05	.01
Verb phrase consisting of "to" and verb phrase	.10	.22	.12	-.22
Verb phrase consisting of base-form verb	-.02	.12	.05	-.04
Verb phrase consisting of base-form verb and adjectival phrase	n/a	n/a	.14	-.04
Verb phrase consisting of past-tense verb and noun phrase	.01	-.15	-.14	-.10

Verb phrase consisting of past-tense verb and subordinate clause	**.06**	**.09**	**.08**	**.30**
Verb phrase consisting of past-tense verb and verb phrase	.19	.11	-.21	.21
Verb phrase consisting of gerund or present-participle	.07	-.13	.07	-.09
Verb phrase consisting of gerund or present-participle verb and noun phrase	.03	-.05	<.01	.08
Verb phrase consisting of gerund or present-participle verb, noun phrase, and prepositional phrase	.17	-.16	.02	-.05
Verb phrase consisting of gerund or present-participle verb and prepositional phrase	.06	<.01	.05	.11
Verb phrase consisting of gerund or present-participle verb and particle phrase	n/a	-.21	n/a	n/a
Verb phrase consisting of gerund or present-participle verb, particle phrase, and noun phrase	-.18	-.06	n/a	n/a
Verb phrase consisting of gerund or present-participle verb, particle phrase, and prepositional phrase	-.12	n/a	n/a	n/a
Verb phrase consisting of gerund or present-participle verb and simple declarative clause	.09	-.07	-.13	.14
Verb phrase consisting of past-participle verb and noun phrase	.05	.03	.12	-.10
Verb phrase consisting of verb and noun phrase	.08	.03	.02	-.14
Verb phrase consisting of past-participle verb and prepositional phrase	.10	.05	-.02	.13
Verb phrase consisting of verb, noun phrase, and prepositional phrase	-.15	-.13	.14	.07
Verb phrase consisting of non-3rd person singular present verb	.22	-.10	.13	-.10
Verb phrase consisting of non-3rd person singular present verb and noun phrase	**-.20**	**-.17**	**-.08**	**-.13**
Verb phrase consisting of verb and prepositional phrase	.04	.29	<.01	.11
Verb phrase consisting of non-3rd person singular present verb and prepositional phrase	.09	-.05	-.02	.09
Verb phrase consisting of non-3rd person singular present verb and simple declarative clause	-.04	.05	-.36	-.23
Verb phrase consisting of non-3rd person singular present verb and subordinating clause	.06	-.25	.04	.04
Verb phrase consisting of non-3rd person singular present verb and verb phrase	.12	-.03	-.09	.08
Verb phrase consisting of a verb followed by a simple declarative clause	-.05	.04	.18	.06
Verb phrase consisting of a verb followed by a subordinating clause	.02	.05	<.01	-.22
Verb phrase consisting of a verb followed by a verb phrase	.14	-.15	.01	-.02
Verb phrase consisting of a 3rd person singular present verb	.12	-.07	-.10	-.14

Verb phrase consisting of a 3rd person singular present verb and an adjectival phrase	-.03	.21	.05	-.07
Verb phrase consisting of a 3rd person singular present verb and a noun phrase	-.01	-.04	.02	.26
Verb phrase consisting of a 3rd person singular present verb, noun phrase, and prepositional phrase	.02	n/a	.12	n/a
Verb phrase consisting of a 3rd person singular present verb and a prepositional phrase	-.05	-.06	-.01	.15
Verb phrase consisting of a 3rd person singular present verb and a simple declarative clause	.07	-.07	-.09	.02
Verb phrase consisting of a 3rd person singular present verb and a subordinating clause	.16	-.01	.08	-.08
Verb phrase consisting of a 3rd person singular present verb and a verb phrase	.12	.04	.15	-.07
Verb phrase consisting of two verb phrases connected with a coordinating conjunction	.05	-.01	.11	<.01
Simple declarative clause consisting of noun phrase and adjectival phrase	**.06**	**.03**	**.05**	**.04**
Simple declarative clause consisting of noun phrase, adverbial phrase, and verb phrase	n/a	n/a	.01	-.04
Simple declarative clause consisting of two noun phrases	<.01	-.27	-.06	-.38
Simple declarative clause consisting of a noun phrase and verb phrase	.16	-.04	.08	-.20
Simple declarative clause consisting of two simple declarative clauses connected by a coordinating conjunction	n/a	n/a	-.15	-.26
Simple declarative clause consisting of a verb phrase	.25	.18	<.01	<.01
Simple declarative clause consisting of a verb phrase followed by a period	n/a	n/a	-.19	n/a
Grammatical complexity				
Clauses	.24	.04	-.40	.08
Complex nominals	-.08	-.03	.09	.29
Complex nominals per clause	-.27	-.06	.27	.15
Complex nominal per T-unit	-.08	.16	.16	.17
Content density: proportion of nouns, verbs, adjectives, and adverbs (Roark)	.19	.18	-.16	.08
Coordinate phrases	**-.16**	**-.07**	**-.01**	**-.13**
Coordinate phrases per clause	-.28	-.08	.03	-.13
Coordinate phrases per T-unit	-.10	.09	.05	-.03
Clauses per sentence	.16	.24	.04	.11
Complex T-units	.04	-.05	-.09	-.22
Complex T-units per T-unit	.01	.23	.01	.06
Dependent clauses	.10	.09	-.04	.04
Dependent clauses per clause	-.01	.03	.06	.05

Dependent clauses per T-unit	<.01	.20	.05	.19
Average length of noun phrases	.07	-.08	.24	.07
Average length of prepositional phrases	-.03	.08	.24	.06
Average length of verb phrases	**.02**	**.09**	**.08**	**.12**
Mean length of clause	-.21	-.10	.27	-.09
Mean length of sentence	**.10**	**.18**	**.09**	**.11**
Mean length of T-unit	-.01	.18	.10	.13
T-unit per sentence	.04	.02	.04	-.09
T-units	**-.01**	**-.07**	**-.28**	**-.03**
Verb phrases per T-unit	**.04**	**.20**	**.01**	**.12**
Length of each noun phrase over total sample length	-.03	-.06	.08	.08
Number of noun phrases over total sample length	-.02	-.12	-.27	.03
Length of each prepositional phrase over total sample length	-.01	.20	.14	.22
Number of prepositional phrases over total sample length	**.06**	**.21**	**.10**	**.20**
Length of each verb phrase over total sample length	.07	.20	-.01	.06
Number of verb phrases over total sample length	.24	.12	-.19	-.2
Average height of each parsed tree in the sample	**.14**	**.15**	**.04**	**.11**
The greatest tree parse depths any all words in the sentence, with weighting for left-branching.	.14	-.14	.30	.06
The sum of tree parse depths for all words in the sentence, with weighting for left-branching	**.15**	**.03**	**.16**	**.11**
The mean of tree parse depths for all words in the sentence, with weighting for left-branching.	.20	-.07	.33	.05
Vocabulary range				
Average length of each word	**.01**	**.05**	**.14**	**.23**
Age of acquisition	.14	.27	.12	-.06
Imageability (subjective rating of how easily a word generates an image in the mind)	**-.21**	**-.03**	**-.05**	**-.29**
Subjective rating of how familiar a word seems	-.06	-.24	-.10	.16
Frequency with which a word occurs in some corpus of natural language	.06	-.29	-.18	-.05
Not-in-dictionary words	**-.11**	**-.11**	**-.31**	**-.11**
Age of acquisition of nouns	.09	.29	.16	-.06
Noun familiarity	-.05	-.23	-.22	.19
Noun frequency	.01	-.10	-.05	.10
Noun imageability	**-.02**	**-.07**	**-.05**	**-.02**
Age of acquisition of verbs	**.10**	**.08**	**<.01**	**.04**
Verb familiarity	<.01	-.10	.13	.08
Verb frequency	**-.07**	**-.11**	**-.18**	**-.15**

Verb imageability	**-.09**	**-.03**	**-.10**	**-.33**
Vocabulary richness				
Brunet's Index	-.04	.09	.34	-.14
Honoré's index	.17	.15	.22	-.04
Moving average type-token-ratio (10-word window)	.12	.09	.24	-.04
Moving average type-token-ratio (20-word window)	.13	.09	.30	-.09
Moving average type-token-ratio (30-word window)	<.01	.10	.27	-.08
Moving average type-token-ratio (40-word window)	.11	.08	.22	-.06
Moving average type-token-ratio (50-word window)	.12	.07	.23	-.05
Type-token ratio	.06	.04	-.33	.20
Word specificity, similarity, and ambiguity				
Average of the averages of each synset's longest WordNet paths to its hypernym/root (all words)	**-.18**	**-.14**	**-.02**	**-.11**
Average of the averages of each synset's longest WordNet paths to its hypernym/root (nouns)	.02	-.12	-.15	.11
Average of the averages of each synset's longest WordNet paths to its hypernym/root (verbs)	.05	-.04	.23	.14
Average of the averages of each synset's shortest WordNet paths to its hypernym/root (all words)	**-.17**	**-.10**	**-.01**	**-.21**
Average of the averages of each synset's shortest WordNet paths to its hypernym/root (nouns)	.06	-.07	-.15	-.09
Average of the averages of each synset's shortest WordNet paths to its hypernym/root (verbs)	.07	-.04	.23	.14
Standard deviation of the longest WordNet paths from given word to hypernym/root	.12	-.02	.04	.05
Standard deviation of the longest WordNet paths from given noun to hypernym/root	-.01	-.07	.13	-.04
Standard deviation of the longest WordNet paths from given verb to hypernym/root	**.27**	**.02**	**.22**	**.03**
Standard deviation of the shortest WordNet paths from given word to hypernym/root	.16	.16	.08	-.10
Standard deviation of the shortest WordNet paths from given noun to hypernym/root	.04	.06	.17	-.11
Standard deviation of the shortest WordNet paths from given verb to hypernym/root	**.27**	**.01**	**.22**	**.03**
Average word meaning similarity (WordNet JCN Brown method)	**-.10**	**-.20**	**-.02**	**-.35**
Average word meaning similarity (WordNet JCN SemCor method)	**-.10**	**-.20**	**-.02**	**-.35**
Average word meaning similarity (WordNet LC method)	-.03	-.05	.05	-.29
Average word meaning similarity (WordNet Lin Brown method)	-.19	-.17	.24	-.34

Average word meaning similarity (WordNet Lin Semcor method)	-.18	-.09	.22	-.42
Average word meaning similarity (WordNet Resnick Brown method)	-.04	.02	-.30	-.02
Average word meaning similarity (WordNet Resnick SemCor method)	**-.04**	**-.20**	**-.26**	**-.14**
Average word meaning similarity (WordNet WP method)	-.16	-.06	.08	-.33
Standard deviation WordNet similarity JCN Brown method	n/a	n/a	.10	-.04
Standard deviation WordNet similarity JCN SemCor method	n/a	n/a	.17	-.11
Standard deviation WordNet similarity LC method	-.09	-.17	.15	-.34
Standard deviation WordNet similarity Lin Brown method	-.08	-.14	.16	-.35
Standard deviation WordNet similarity Lin SemCor method	-.14	-.13	.14	-.37
Standard deviation WordNet similarity Resnick Brown method	.16	-.26	.28	-.16
Standard deviation WordNet similarity Resnick SemCor method	.17	-.19	.22	-.19
Standard deviation WordNet similarity WP method	-.17	-.10	.21	-.28
Average WordNet ambiguity (all words)	-.09	-.12	-.09	.05
Average WordNet ambiguity (nouns)	-.01	.17	<.01	.08
Average WordNet ambiguity (verbs)	-.18	-.11	.08	<.01
Kurtosis WordNet ambiguity (all words)	<.01	.18	.19	-.15
Kurtosis WordNet ambiguity (nouns)	.01	.20	.26	.02
Kurtosis WordNet ambiguity (verbs)	-.01	.21	.28	-.15
Skewness WordNet ambiguity (all words)	.01	.14	.23	-.09
Skewness WordNet ambiguity (nouns)	.06	.09	.15	.01
Skewness WordNet ambiguity (verbs)	.11	.25	.23	-.18
Standard deviation WordNet ambiguity (all words)	-.02	-.05	.14	.03
Standard deviation WordNet ambiguity (nouns)	.14	.17	.29	<.01
Standard deviation WordNet ambiguity (verbs)	.16	.06	.22	.06
Vocabulary sentiment				
Mean Stanford Sentiment Negative	.13	-.05	.15	-.08
Mean Stanford Sentiment Neutral	-.02	.11	-.21	.25
Mean Stanford Sentiment Positive	-.19	.15	-.12	-.04
Mean Stanford Sentiment Very negative	-.02	-.04	.22	-.10
Mean Stanford Sentiment Very positive	**-.16**	**-.05**	**-.03**	**-.02**
MPQA Strong negative	**.20**	**.02**	**.01**	**.09**
MPQA Strong positive	.18	-.07	.10	.21
MPQA Weak negative	-.10	.06	.05	.01

MPQA Weak positive	-.24	.02	.03	-.31
Arousal mean	.08	-.21	.06	.12
Arousal mean nouns	.03	-.29	.01	.21
Arousal mean verbs	.06	-.18	-.04	.08
Arousal standard deviation	.12	-.22	-.09	-.13
Arousal standard deviation nouns	.11	-.24	-.08	-.09
Arousal standard deviation verbs	.14	-.16	-.04	-.05
Dominance mean	.05	<.01	.04	.15
Dominance mean nouns	.01	-.07	.08	-.07
Dominance mean verbs	-.01	.21	.04	.22
Dominance standard deviation	.02	-.18	-.17	.16
Dominance standard deviation nouns	.19	-.15	-.07	.33
Dominance standard deviation verbs	-.05	-.16	-.10	.04
Valence mean	-.04	-.06	-.18	.14
Valence mean nouns	-.01	-.14	-.07	.04
Valence mean verbs	-.06	.18	-.15	.18
Valence standard deviation	-.08	-.16	.25	.12
Valence standard deviation nouns	-.01	-.20	.03	.07
Valence standard deviation verbs	-.10	-.06	.23	.01

Note: Features with "n/a" instead of a correlation result means that feature was not present in the data or was removed from a given dataset during preprocessing due to zero or near-zero variance. When the direction of correlation between a feature and reading comprehension is the same across all datasets, that feature is bolded.
[a]Function words include determiners, personal pronouns, possessive pronouns, wh-determiners, wh-pronouns, possessive wh-pronouns, coordinating conjunctions, particles, modals, preposition, and subordinating conjunctions

Appendix B

Table 35

BALA-Regular (N=132) item-level statistics

Item number	Facility (P)	Discrimination (Pearson point-biserial)
1	.89	.35
2	.87	.43
3	.59	.20
4	.83	.35
5	.73	.17
6	.83	.43
7	.58	.19
8	.36	.27
9	.37	-.01[1]
10	.67	.42
11	.68	.62
12	.80	.55
13	.36	.37
14	.78	.72
15	.48	.29
16	.67	.59
17	.70	.54
18	.67	.52

[1] The reliability of the BALA-Regular measure was checked without this item included. The increase in reliability was negligible (increase of less than .02). Therefore, the item was retained.

Table 36

BALA-Modified (N=109) item-level statistics

Item number	Facility (P)	Discrimination (Pearson point-biserial)
1	.62	.59
2	.76	.60
3	.81	.62
4	.68	.58
5	.79	.62
6	.63	.47
7	.73	.58
8	.80	.64
9	.73	.65
10	.62	.31
11	.68	.45
12	.79	.42
13	.69	.48
14	.57	.15
15	.10	.22
16	.65	.37
17	.66	.18