COGNITION, MODALITY, AND LANGUAGE
IN HEALTHY YOUNG ADULTS

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

---

by
Ann Marie Finley
December 2023

Examining Committee Members:

Lisa Bedore, PhD, Advisory Chair, Department of Communication Sciences and Disorders
Jodi Reich, PhD, Department of Communication Sciences and Disorders
Rena Krakow, PhD, Department of Communication Sciences and Disorders
Nadine Martin, PhD, Department of Communication Sciences and Disorders
Yuexiao Dong, PhD, Department of Statistics, Operations and Data Science
Manaswita Dutta, PhD, External Member, Portland State University

ABSTRACT

Measures drawn from language samples (e.g., discourse measures) are used in clinical and research settings as a functional measure of language and cognitive abilities. In narrative elicitation tasks, discourse measures reliably vary by the type of prompt used to collect a language sample. Additionally, language features tend to very along with communicative context, topic, and modality (e.g., oral vs. written). However, until recent years, technology had not advanced sufficiently to support large-scale study of spoken language data. In this project, we used natural language processing and machine learning methods to examine the intersection of discourse measures, language modality, and cognition (i.e., working memory) in healthy young adults. In Experiment 1, we used a computational approach to examine discourse measures in spoken and written English. We achieved >90% accuracy in binary classification (e.g., spoken/written). In Experiment 2, we took a behavioral approach, studying working memory and narrative discourse measures in a cohort of healthy young adults. We predicted that working memory would predict informativity in participants' narrative language samples. We found mixed results for our two measures of *informativity* (e.g., the Measure of Textual Lexical Diversity and Shannon entropy). We attributed the observed differences in these two measures to the fact that, while both serve to measure new or unique information, MTLD indexes additional linguistic information (e.g., semantic, lexical). In contrast, Shannon entropy is based on word co-occurrence statistics. We interpret our overall results as support for the potential utility of machine learning in language research and potential for future research and clinical implementations.

Jointly dedicated to

Elizabeth Idell Kelly, 1990 – 2021

and

All researchers, clinicians, and scholars who have persisted in their pursuit of excellence
when faced with the scope and depth of the racism, sexism,
and other forms of discrimination that historically
and systemically plague institutions of power
within modern culture.

ACKNOWLEDGEMENTS

A large debt of gratitude is owed to the mentors, colleagues, and friends that have supported my efforts to become a better-informed, more critically engaged member of society. To Frank Franco, Claire Nouvet, Susan Tamasi, Laura Namy, Edie Hapner, Lisa LaGorio, and Emily Wang: Thank you for the time, care and attention you took in providing me with educational and research experiences that inspired me to return to school in pursuit of higher education again and again, despite all of my best intentions to avoid spending over five years in grad school. As I have learned over the course of my doctoral studies, completed alongside an invaluable pair of colleagues and my fellow Fall 2019 cohort-mates Matt Sayers and Maura O'Fallon, time is fake and math is everything. Much appreciation goes out to Jodi Reich, Gayle Dede, and Lisa Bedore for their efforts to support my research through a period of significant adversity faced in the context of a global pandemic and public health emergency. Finally, to my advisory and examining committee members: Lisa Bedore, advisory chair; Jodi Reich, examining chair; Nadine Martin; Rena Krakow; Yuexiao Dong; and outside examiner Manaswita Dutta – thank you for the knowledge, laughter, discussions, and enthusiasm each of you showed along my research journey. In particular I'd like to thank these individual sources of inspiration: Jen B., Becca B., Meg S., Celia L., Rachel O., Jill K., Bonnie Z., and Athena R., to name just a few. Most of all, thanks to my parents, Davis and Nancy; my siblings; my niblings (i.e., nieces and nephews); and my entire extended family for their love, support, and comic relief provided over the past five years.

TABLE OF CONTENTS

Page

CHAPTER

APPENDICES

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

GENERAL INTRODUCTION

A growing area of interest in the evaluation and treatment of cognitive-linguistic disorders involves supplementing standard neuropsychological testing with the collection and analysis of naturalistic language samples (e.g., storytelling, conversation; Bryant et al., 2016; Fraser et al., 2015; MacWhinney et al., 2011; Mota et al., 2017; Orimaye et al., 2017; Stark, 2019; Stark et al., 2022). Features derived from naturalistic language samples (i.e., discourse measures) range from the micro-level (e.g., psycholinguistic indices such as word-level *concreteness* or *frequency*) to the macro-level (e.g., syntactic complexity, story grammar; Bryant et al., 2016; Kong et al., 2016; Stark et al., 2022). Across a range of disciplines, researchers have documented a relationship between indices of cognitive-linguistic function (e.g., scores on standardized assessments, clinical diagnosis of disease) and various discourse measures. For example, people with mild cognitive impairment (MCI) tend to produce less syntactically complex stories compared to healthy controls (Roark et al., 2011). Analysis of student-written essays reveals that students who use more lexically sophisticated vocabulary (i.e., less frequently-occurring words) tend to achieve better academic outcomes relative to peers who use less sophisticated vocabulary (Crossley, 2020; Crossley & Allen, 2016; McNamara et al., 2010).

Informativity is a construct that conceptualizes the quantity of new or unique information in a given language sample (Shannon, 1950; Shannon & Weaver, 1949). Over the years, various discourse measures have been proposed and implemented to index the *informativity* of a given language sample (K.T. Cunningham & Haley, 2020; Florian Jaeger, 2010; Fraser et al., 2015; Garrard et al., 2014; Garrard & Forsyth, 2010; Mitzner & Kemper, 2003; Nicholas & Brookshire, 1993; Shannon, 1950; Shannon & Weaver, 1949; Sirts et al., 2017; Yancheva & Rudzicz, 2016). These informativity measures tend to vary within and across research disciplines; however, by surveying the overall pattern of results yielded by the study of informativity within language, some general interpretations may be drawn about the relationship between cognitive-linguistic function, informativity, and the effect of language modality on these measures. These interpretations, in turn, lead to questions regarding the interrelationship of informativity, cognitive processing, and language modality. It is such questions that this paper aims to address. To follow, we present an overview of the theoretical and empirical motivations driving the current investigation.

Informativity, Frequency, and Redundancy

Computer scientists, computational linguists, and neuroscientists performing discourse analysis have consistently documented a relationship between various measures of informativity (e.g., correct information units, count ratio of nouns to verbs) and presence or risk of dementia among older adults (Fraser et al., 2015; Garrard et al., 2014; Garrard & Forsyth, 2010; Sirts et al., 2017; Yancheva & Rudzicz, 2016). Notably, early work examining informativity in language was largely agnostic to word-level lexical and phonological features, focusing rather on frequency of occurrence for a given lexical-semantic representation (Shannon, 1950; Shannon & Weaver, 1949). Depending on the modality of language production, lexical-semantic representations (i.e., words) may take the form of a sequence of sounds (e.g., phonology as in spoken language) or a visual orthography (e.g., as in written language). Word frequency and informativity have been linked through their respective relationships to the concept of redundancy within language. Across word types, those that occur more frequently may also be characterized as more *redundant* than words that occur less frequently – that is, repeated exposure to the same word (i.e., the same set of phonological or orthographic symbols) yields greater familiarity with said word (Shannon, 1950; Shannon & Weaver, 1949). The more familiar a word, the less surprising or informative it is considered to be; e.g., the greater *redundancy* it carries. More frequently-occurring (e.g., more redundant) words tend to be recalled more easily and produced more efficiently than less frequently-occurring words (Florian Jaeger, 2010; Monsell et al., 1989; Nicholas & Brookshire, 1993). Similarly, words that tend to frequently co-occur may be compressed in order to increase efficiency at the phrase, paragraph, or discourse level.

For example, shortening the phrase *going to* → *gonna* reduces the total number of syllables from three to two while maintaining the semantic information contained in the symbolic representation in question (Florian Jaeger, 2010; Lewis & Frank, 2016; Nicholas & Brookshire, 1993). Interpreted within Shannon's information-theoretic approach to language, one might argue that this phonemic and syllabic compression serves to increase the informativity of the phrase by maximizing the efficiency of the signal (Shannon, 1950; Shannon & Weaver, 1949). That is, by decreasing the number of syllables and/or phonemes required to convey the same conceptual-semantic meaning, each remaining unit of language carries a greater degree of information. In this example, by conveying an equivalent amount of information via a more compressed signal, the *information density* (i.e., efficiency) of the message is increased (Shannon, 1950;

Shannon & Weaver, 1949). The construct of *information density* is particularly interesting when considered within theories of spoken and written language production.

Language Modality: Spoken vs. Written

Observing a greater number of non-content filler words (e.g., *um, uh)* in spoken vs. written language samples, researchers in the past theorized that individuals were subconsciously decreasing their rate of information transmission in order to attenuate listeners' working memory demands associated with spoken vs. written forms of expression (Basso et al., 1978; Biber, 2004; Chafe & Tannen, 1987; Fergadiotis & Wright, 2011; Mitzner & Kemper, 2003). In other words, they proposed that adding non-content filler words (e.g., *um, uh*) to a language sample decreased information density at the sample level by reducing word-level informativity. One way to conceptualize the putative link between working memory and information density is through a cost reduction framework, where cost represents the relative demands enacted on working memory in language processing.

Recall from our *going to* → *gonna* example that as the number of language units *s* decreased, information density increased. The process of language production ostensibly entails some degree of cognitive effort on the part of the author (Ben Shalom & Poeppel, 2008; Dell & Anderson, 2015; Flower & Hayes, 1981; Kellogg et al., 2013; Piantadosi et al., 2011). Psychology and evolutionary biology tell us that the human brain is optimized for efficiency; thus, by decreasing the number of language units *s* required to convey the intended meaning, the effort (i.e., cost) of language production is decreased (Florian Jaeger, 2010; Piantadosi et al., 2011). Put simply, if language unit *s* no longer appears in a given language sample, then the cost associated with producing *s* is also eliminated. If the overall informativity of the language sample is maintained during signal compression, as in *going to* → *gonna*, then it is a mathematical necessity that, as the number of language units *s,* represented in theoretical vector *S*, decreases by an integer value *n,* information density increases in the remaining sample *s* – *n*. On the surface, this reasoning seemingly contradicts Shannon's information-theoretic framework (Shannon, 1950; Shannon & Weaver, 1949). Briefly, Shannon asserts that increasing the number of words (i.e., language units *s*) in language sample *LS* is analogous to increasing the degrees of freedom in a mathematical model, thereby increasing entropy, i.e., the amount of possible variance within a given system. Thus, as word count increases, so too do degrees of freedom, accompanied by greater variability and greater uncertainty as to the source of the variability (Shannon, 1950; Shannon & Weaver, 1949). However, it is well established that human language processing is impacted by a wide range of lexical, phonemic, and semantic features (Chafe & Tannen, 1987; Piantadosi et al., 2011). Indeed, this reasoning holds up when examined in light of researchers' earlier claim that increasing the number of language units *s* (i.e., words) decreases working memory demands in spoken language only if overall *informativity* remains constant (Basso et al., 1978; Biber, 2004; Chafe & Tannen, 1987; Fergadiotis & Wright, 2011; Mitzner & Kemper, 2003). Inserting non-meaningful words (e.g., *um,* uh) yields a reduction in the average informativity of a given word *s* within language sample *LS*. It is not illogical to assume that such a decrease in word-level informativity corresponds to a similar decrease in the overall informativity of language sample *LS.* Further, given that spoken and written language inherently differ in their respective manner and mode of production, it is likely that the cost associated with increasing information density varies along with language modality. While both spoken and written theories of language production cite working memory as a critical process component, additional parameters, including pragmatic and contextual information, must be considered when characterizing the relationship of language modality, working memory, and information density in language processing. To follow, we provide a brief overview of critical pragmatic and contextual factors thought to influence spoken and written expression.

The Influence of Audience: Pragmatics and Context

In the absence of sensorimotor degradation or deprivation, humans experience spoken language primarily through audition, recruiting auditory processing pathways to decode meaning from sound sequences. Speakers commonly manipulate vocal features (e.g., prosody, pitch) to convey affective information and signal non-literal language use (e.g., sarcasm, irony; see Richter & Chatterjee, 2021). Across many naturally-occurring spoken language contexts, facial expressions, gestures, and bodily orientation serve as conscious and unconscious cues to the successful interpretation of a speaker's intended meaning. In these interactive contexts, pragmatically-induced temporal constraints limit opportunities for extended message planning, potentially leading to errors in spoken expression (e.g., semantic errors, spoonerisms). Although the aurality of spoken language precludes opportunities for revision once a message is produced (Auer, 2009), the dynamic nature of spoken communication allows for real-time listener feedback to flag and subsequently repair communicative failures.

In contrast, written expression occurs in a context removed from the message receiver. In the absence of paralinguistic cues to signal emotion or non-literal language use, written expression relies solely on lexical-semantic (Richter & Chatterjee, 2021) and syntactic (Whalen et al., 2013) features to convey information. In informal written language contexts (e.g., email, social media), research shows that manipulating punctuation (e.g., *Thanks!* vs. *Thanks,*) and spelling (e.g., *heyyy* vs. *hey*) can be effective ways to convey positive affect and other pragmatic information (Darics, 2013; Marlow et al., 2018). Writers must anticipate and address the needs of an imagined audience in order to successfully convey their intended meaning (Fussell & Krauss, 1992; Sauerland & Sporer, 2011). This process was termed "thinking for writing" by linguist Daniel Slobin (2018). Thinking for writing is thought to involve planning and revision throughout the composition of a written text (Epting et al., 2013; Flower & Hayes, 1980, 1981). Free of the temporal constraints of spoken expression, writers are able to spend time crafting and refining the structure and content of their work to accurately represent their intended meaning. In its finished form, written language mirrors the linearity of spoken expression (Auer, 2009). However, the different pragmatic and contextual demands of written vs. spoken language production are thought to influence observed differences in the syntactic, lexical, and semantic features characteristic of each modality (Biber, 1986, 2004; Chafe & Tannen, 1987). In the current project, we focused in particular on *information density* as it relates to working memory in spoken vs. written language.

Information Density, Working Memory, and Language Modality

Information density is a construct that refers to the dispersion of information across a language sample. It is measured in different ways across the various professions engaged in the study of language (K.T. Cunningham & Haley, 2020; McCarthy & Jarvis, 2010; Shannon, 1950) and generally observed to vary such that *information density*$_W$ >

information densitys, a finding attributed to differential working memory demands enacted in oral vs. written expression. However, information density, along with other discourse measures, is known to vary by language task (e.g., “Tell me a story” vs. a conversational exchange), along with contextual factors (e.g., pragmatic demands at work vs. at home). Thus, it is unclear whether modality contributes independent variability to measures of information density across different contexts and genres. By extension, it is unclear whether these measures, previously tied to working memory, do in fact reflect a differential working memory demand enacted in oral vs. written language expression. Here we present two studies that set the stage for a productive research program investigating the relationship of information density, working memory, and language modality in naturalistic language samples. Using both computational and behavioral methods, paired with appropriate use of advanced statistical analyses, we explore the relationship between discourse measures, cognitive systems and functioning, and language modality (i.e., spoken vs. written English) in two experiments.

Experimental Overview

In Experiment One, we use natural language processing methods to analyze the effect of modality on language measures indexing *informativity* (i.e., *idea density*), along with psycholinguistic variables known to drive variation in performance on single-word processing tasks (e.g., lexical decision time). We aimed to determine what, if any, effect modality has on discourse measures over and above known effects of genre and context. We predicted that a machine learning classification algorithm trained on 80% of the total language data would reliably and accurately classify the remaining 20% by modality using only discourse measures.

In Experiment Two, we work from a simple model of language production encompassing both spoken and written expression, with a focus on the role of working memory as a predictor of information density and lexical diversity (i.e., *informativity* or *idea density*) in narrative language samples elicited from healthy young adults. We predicted that working memory positively predicts information density and lexical diversity (i.e., *informativity/idea density*) in narrative language samples; with stronger effects in spoken than in written language.

CHAPTER 2

EXPERIMENT ONE

Abstract

Purpose: We aimed to determine what, if any, effect modality has on discourse measures over and above known effects of genre and context. We predicted that a machine learning classification algorithm will reliably and accurately classify texts by modality using only discourse measures. **Method:** We chose the Spotify Podcast Dataset, containing 622,115,467 words, as a representative corpus of naturalistic spoken English. Written subsections of the Corpus of Contemporary American English (i.e., all genres but spoken and TV/movies) represented the comparison corpus of written English, totaling 641,410,953 words. After applying a cleaning and pre-processing pipeline to the data using the R programming language, we extracted discourse measures linked to human language processing. We trained a machine learning classification algorithm on 80% of the resultant data across modalities. We then tested whether the algorithm was capable of correctly classifying the remaining 20% of texts into either the spoken or written modality. **Results:** Across our full set of discourse measures and a series of subsequent leave-one-out analyses, we reached 93.15% accuracy in spoken/written text classification using a support vector machine classifier. **Conclusions:** We found support for our prediction that spoken and written language are statistically discriminable based on a curated set of discourse measures. In the future, it will be important to expand this research to examine interaction effects of modality and communicative context on our chosen discourse measures.

Information Density and Lexical-Semantic Features of Spoken vs. Written English

Introduction

Over many decades of research, the question of how certain features of language differ in speaking vs. writing has been addressed by linguists (Biber, 1986; Chafe & Tannen, 1987), psychologists (Cleland & Pickering, 2006; Louwerse et al., 2004; Mitzner & Kemper, 2003), cognitive neuroscientists (Brownsett & Wise, 2010; Planton et al., 2013), and speech-language pathologists. Of particular interest to the current investigation are two discourse measures linked to working memory (WM) and long observed to vary between spoken and written language: information density (i.e., how information is distributed across a language sample) and lexical diversity (Biber, 1986; Blankenship, 1974; Chafe & Tannen, 1987; Mitzner & Kemper, 2003; Rubin, 1987; Slobin, 1997). These two measures each serve to represent an overarching linguistic construct that may be broadly termed *idea density* or *informativity* – that is, how tightly packed with fresh or less predictable information a given text is. Predictive modeling indicates that spoken language idea density can be used to discriminate neurologically intact older adults from those with Alzheimer's disease or its sometimes-precursor, mild cognitive impairment (Sirts et al., 2017; Yancheva & Rudzicz, 2016). Similarly, students who use more *lexically sophisticated* vocabulary (e.g., lower frequency, lower concreteness) in essay writing tend to achieve better academic outcomes relative to peers who tend to use less lexically sophisticated words (Crossley, 2020). Altogether, it seems that language features may serve as a useful biomarker of neurocognitive function and disease potential. Yet, absent the knowledge of what drives reported differences in features of spoken and written language (e.g., modality vs. context or other factors), it is difficult to draw conclusions about the cognitive systems putatively supporting language production in speaking vs. writing. In the current project we explore this question using large and representative corpora of English. In doing so, we aim to isolate effects of modality from other variables known to influence language processing (e.g., individual vocabulary knowledge, social setting) with downstream effects on related discourse measures.

Information Density and Lexical Diversity in Spoken vs. Written English

Converging evidence from a variety of disciplines supports a relationship between working memory and language processing in speaking (Baddeley, 2003; Cowan, 1999; Dell & Anderson, 2015; Gilchrist et al., 2008; Martin et al., 2018, 2020) and writing (Crossley, 2020; Crossley & Allen, 2016; Kellogg et al., 2016; McNamara et al., 2010; Olive & Kellogg, 2002; Sauerland et al., 2014). Two discourse-based language measures linked to working memory are information density and lexical diversity. Decades of cross-disciplinary research indicate that information density and lexical diversity tend to vary by modality (Basso et al., 1978; Biber, 2004; Chafe & Tannen, 1987; Fergadiotis & Wright, 2011; Mitzner & Kemper, 2003). For example, one measure of information density (i.e., propositional density), indexes the amount of information conveyed in a text (i.e., informativeness) using a count ratio of propositions to total number of words (Turner & Green, 1977). Evidence from participants in the Nun Study (Snowdon et al., 1996) shows that intra-individual information density is greater in written than oral narratives, such that mean information conveyed per word is greater in writing vs. speaking (Mitzner & Kemper, 2003). Measures of lexical diversity (e.g., type-token ratio, TTR) tend to be smaller in oral vs. written expression, indicating more restricted vocabulary use in oral expression (Louwerse et al., 2004; Mitzner & Kemper, 2003). Language researchers have attributed these observed differences in information density and lexical diversity to differential working memory (WM) demands enacted in speaking versus writing. However, absent the knowledge of modality effects on lexical diversity and information density across a broad range of communicative contexts, it is difficult to interpret the theoretical significance of the apparent link between lexical diversity, information density, and working memory as it pertains to differential cognitive demands enacted in speaking vs. writing. Thus, it is critical to determine what, if any, effects modality has on information density and lexical diversity.

Psycholinguistic Features of Spoken and Written English

There is general agreement that written language tends to occur in more formal (e.g., newspapers, legal briefs) contexts and therefore elicits use of a more sophisticated register than spoken language (Blankenship, 1974; Chafe & Tannen, 1987; Louwerse et al., 2004). It is well-established that psycholinguistic variables (e.g., concreteness, frequency) influence language processing at the single-word level (Balota et al., 2007; Pexman et al., 2017). Many psycholinguistic variables are intercorrelated: words that are longer tend to be acquired later (Kuperman et al., 2012) and more often refer to abstract vs. concrete concepts (Brysbaert et al., 2014). In turn, earlier-acquired words tend to be shorter, occur more frequently, and refer to concrete concepts (Brysbaert & Ghyselinck, 2006).

Research in cognitive language science indicates a positive correlation between lexical sophistication and response latency on single-word processing tasks (Balota et al., 2007; Pexman et al., 2017). This suggests that on average, neurotypical adults are slower to process more lexically sophisticated words, although individual differences may be observed (Pexman & Yap, 2018; Yap et al., 2012). To our knowledge, there is currently no published research broadly characterizing lexical sophistication in oral vs. written expression. It is possible that spoken language is less lexically sophisticated than written language secondary to relative contextual formality. This idea converges with theories on WM load reduction strategies in oral vs. written expression (Chafe & Tannen, 1987; Slobin, 1997). Just as it is possible that speakers lessen WM load by using a restricted range of vocabulary words and conveying information more slowly (e.g., decreased *idea density*), it is possible that using less lexically sophisticated vocabulary further alleviates WM demands associated with language production. In the current study, we examine *lexical sophistication*, along with *idea density*, measured by information density and lexical diversity, in predicting text modality (e.g., spoken vs. written) using a form of advanced statistical analysis broadly referred to as machine learning.

Machine Learning in Language Research

Until recent decades, researchers lacked the computational and methodological tools to analyze the effect of modality (e.g., speaking vs. writing) on discourse measures across a broad range of communicative contexts. This reality reflects a broader perspective: knowledge of the world is constrained by the availability and quality of the technologies used in its measurement and analysis. Even after the invention of the tape recorder allowed for the collection of spontaneous, naturalistic spoken language data, researchers faced hours of manual transcription necessitated by speech's innate ephemerality (Chafe & Tannen, 1987; Edwards & Lampert, 1993). More recently, advances in speech-to-text technology have enabled researchers to at least partially outsource transcription to automated software systems, thus further alleviating the arguably greater demands engendered in researching spoken vs. written language (Chafe & Tannen, 1987; Clifton et al., 2020; Goh et al., 2020). Accompanying advances in computing and statistics have since intersected to facilitate language researchers' ability to create, store, and apply quantitative analysis to large datasets representing spoken (Clifton et al., 2020) and/or written language. Traditional statistical analyses (e.g., ANOVA, linear regression) are insufficient to fully characterize language features across large corpora of spoken and written texts. Rather, as computing power and available data continued to expand, researchers developed new statistical methods targeted at extracting meaningful information from massive datasets. The resultant field of study, known as machine learning, has yielded widely-used applications including financial and political forecasting tools, facial identification software, and email spam filters (Lantz, 2013). Email spam filters are a representative application of *supervised* machine learning. In supervised learning, a machine learning algorithm is applied to model relationships in the data that contribute to a specific outcome (e.g., spam email vs. not spam email). Typically, this is accomplished by splitting data in to a training set and a relatively smaller testing set. The training set is used to develop the output classification model, which is then assessed using the testing set. We review one widely-used and well-established approach to supervised machine learning in the following section.

Support Vector Machine Classifiers

Support vector machines (SVMs) were developed in the late twentieth century with the goal of improving binary classification tasks (Berwick, 2003; Cortes & Vapnik, 1995). Today, SVMs are a well-established approach to multiple-group classification and regression tasks, applied in a range of research disciplines (Bennett & Campbell, 2000; Lantz, 2013; Meyer, 2023). Although SVMs are considered a “black box” technology due to the sheer complexity of the algorithms used to build the classification model, the underlying mathematical operations have existed for decades (Bennett & Campbell, 2000; Berwick, 2003; Lantz, 2013). In essence, SVM algorithms take an array of vectors as input and perform non-linear mapping into a high dimensional feature space. Advanced calculus is then used to determine the best-fit line separating two (or more) categories (Bennett & Campbell, 2000; Cortes & Vapnik, 1995; Lantz, 2013). Rather than weighting each data point equally in determining the best-fit line, as in linear regression, SVMs make use of ‘support vectors’ to optimize category margins. ‘Support vectors’ are the data points closest to category boundaries and are critical pieces of the optimization algorithm calculating the best-fit line (called a ‘hyperplane’ in feature spaces greater than two dimensions). Through iterative testing, SVMs optimize the hyperplane to achieve the widest possible confidence interval between groups (Lantz, 2013; Meyer, 2023). SVMs tend to generate highly accurate models with good generalization to new data (Bennett & Campbell, 2000; Berwick, 2003; Cortes & Vapnik, 1995; Lantz, 2013). Applied to language corpora, such approaches can help us to understand what discourse measures reliably vary between spoken and written expression by extracting and analyzing patterns of relationships not detectable to the human eye.

The Present Study

A combination of technological and computational advances in recent decades has facilitated researchers’ ability to collect, transcribe, and analyze naturalistic spoken language data. In the current study, we contrast features of spoken and written language with a particular focus on lexical diversity and information density. We predict that across a wide range of contexts and genres, spoken and written language are discriminable based on statistical distributions of discourse features and psycholinguistic measures drawn from naturalistic language samples. Our primary goal in this investigation was to determine whether information density and lexical diversity consistently differ by modality in their overall distributions across many contexts and genres of language use. We further explored whether spoken and written language systematically vary along various psycholinguistic indices linked to human language processing.

Methods

To address our predictions, we used natural language processing methods to extract all variables of interest from large and representative corpora of spoken and written English. We then used a support vector machine algorithm to determine whether discourse measures and psycholinguistic features can successfully predict text modality (e.g., spoken vs. written).

Spoken English Corpus: The Spotify Podcast Dataset

We selected the 622,115,467-word Spotify Podcast Dataset (Clifton et al., 2020) to represent our corpus of spoken English. The Spotify Podcast Dataset contains audio files and transcriptions for 18,376 podcasts (n = 105,360 unique episodes) released on the Spotify platform between January 1, 2019 and March 1, 2020. Podcasts were randomly sampled and cover a wide range of subjects (e.g., science, pop culture) and discourse styles (e.g., conversational, technical). Approximately 10% of podcasts included in the dataset were professionally produced, while the remaining 90% came from amateur creators. Our goal in corpus selection was to provide a comprehensive depiction of contemporary spoken English across a variety of contexts and genres. The Spotify Podcast Dataset meets these criteria; however, it is not a perfect representation of spontaneously produced spoken language. Portions of the included podcasts include scripted material (e.g., introductions) and episodes were likely edited prior to publication on the Spotify platform. To address the validity of the Spotify Podcast Dataset as a representation of naturalistic spoken language, we compared the Spotify Podcast Dataset with various other corpora of spoken English (see Table 1).

Table 1. Comparison corpora of spoken English

Corpus		**Pub. Year**	**Genre**	**Documents**	**Tokens**
Spotify Podcast Dataset		2021	Varied	N=105,360 (n=5,268)	622,115,467
Buckeye Corpus		2000	Conversation	248	249,000
Open American National Corpus	*Charlotte*	2002	Conversation; narratives; interviews	93	198,295
	Switchboard	1992	Conversation	2,307	3,019,477

Note. Pub. = publication. Charlotte = UNC Charlotte Narrative and Conversation Collection. Switchboard = Switchboard Corpus at UPenn, (Open American National Corpus, 2015). Buckeye = Iowa Buckeye Corpus, (Pitt et al., 2007).

All of the comparison corpora included conversational spoken language; while one corpus additionally included narratives and interviews. Table 2 provides an overview of summary statistics for the four compared corpora.

Table 2. Corpora features comparison overview

	Spotify Podcast Dataset (N=5254)	Buckeye (N=248)	Charlotte (N=93)	Switchboard (N=2307)
Word Count				
Mean (SD)	5930 (4320)	1260 (434)	2140 (1670)	1330 (488)
Median [Min, Max]	5270 [53.0, 32500]	1300 [115, 2240]	1880 [252, 10900]	1130 [126, 3100]
Syllables				
Mean (SD)	1.30 (0.0868)	1.26 (0.0414)	1.24 (0.0463)	1.23 (0.0440)
Median [Min, Max]	1.28 [1.03, 1.87]	1.25 [1.17, 1.44]	1.24 [1.14, 1.35]	1.22 [1.12, 1.40]
Letters				
Mean (SD)	4.01 (0.265)	3.82 (0.129)	3.82 (0.144)	3.71 (0.137)
Median [Min, Max]	3.94 [3.21, 5.51]	3.82 [3.49, 4.19]	3.82 [3.44, 4.19]	3.70 [3.30, 4.25]
Concreteness				
Mean (SD)	2.50 (0.0999)	2.50 (0.0774)	2.56 (0.0871)	2.47 (0.0670)
Median [Min, Max]	2.49 [2.21, 4.66]	2.50 [2.31, 2.73]	2.57 [2.34, 2.74]	2.47 [2.23, 2.69]
Age of Acquisition				
Mean (SD)	5.05 (0.275)	4.90 (0.150)	4.70 (0.140)	4.90 (0.160)
Median [Min, Max]	4.98 [4.14, 7.01]	4.89 [4.56, 5.47]	4.70 [4.39, 5.09]	4.87 [4.51, 5.48]
Phonemes				
Mean (SD)	3.18 (0.233)	3.01 (0.117)	3.01 (0.128)	2.93 (0.120)
Median [Min, Max]	3.12 [2.57, 4.73]	3.00 [2.71, 3.44]	3.02 [2.66, 3.30]	2.92 [2.60, 3.44]
Frequency				
Mean (SD)	7350 (673)	6690 (553)	7200 (634)	7350 (514)
Median [Min, Max]	7380 [687, 11100]	6660 [5250, 8610]	7200 [5650, 9520]	7330 [5270, 9540]

	Spotify Podcast Dataset (N=5254)	**Buckeye (N=248)**	**Charlotte (N=93)**	**Switchboard (N=2307)**
MTLD				
Mean (SD)	58.5 (20.1)	44.5 (9.52)	46.6 (9.03)	41.7 (7.56)
Median [Min, Max]	53.2 [3.36, 256]	45.1 [22.2, 74.6]	45.4 [28.0, 71.5]	41.0 [21.6, 83.6]
Shannon entropy				
Mean (SD)	5.43 (0.344)	5.00 (0.200)	5.12 (0.232)	4.95 (0.134)
Median [Min, Max]	5.48 [1.45, 6.55]	5.04 [4.15, 5.39]	5.13 [4.37, 5.67]	4.95 [4.13, 5.36]

Note. SD = standard deviation; Min = minimum; Max = maximum. Word count = mean word count per document; MTLD = Measure of Textual Lexical Diversity; Syllables = mean number of syllables per word; Letters = mean word length in letters; Concreteness = mean per word, (Brysbaert et al., 2014); Age of Acquisition = mean per word, (Kuperman et al., 2012); Phonemes = mean per word; Frequency = word frequency, indexed as mean count per million words, from SUBTLEX-US (Brysbaert & New, 2009).

Visual inspection reveals that Spotify has a slightly higher mean than the other three corpora in word count, followed by Charlotte, then Switchboard and Spotify. This pattern of results is not unexpected since the Charlotte corpus includes narrative language samples in addition to the conversational language samples also present in the other comparison corpora. Narrative language (e.g., personal or fictional stories) may be used in conveying more complex sequences of chronological or non-chronological events than conversational language, thereby necessitating the use of a greater number of words in narrative language. By extension, it is unsurprising that the mean word count per document in the Spotify Podcast Corpus (comprised of a range of genres) is over twice that of the Charlotte corpus. Spotify and Charlotte both have greater range and variance in their respective word count distributions when compared to the conversation-only spoken language corpora. However, along several psycholinguistic variables, the mean values for the four corpora tend to be relatively close in value. In general, variables measured from Spotify tend to contain higher upper bounds compared to the other three corpora. This was more observable in some variables (e.g., MTLD) than others (e.g., mean syllables per word). Here, it is again possible that we are seeing an effect of the wide variety of genres included in the Spotify Podcast Dataset. On the whole, the distribution of variables in the Spotify Podcast Dataset appears similar to those drawn from the comparison corpora, particularly for word-level variables (e.g., mean concreteness) that may be less affected by modality than higher-level variables, such as MTLD (i.e., lexical diversity) and Shannon entropy (i.e., information density).

Written English Corpus: The Corpus of Contemporary American English

To represent written English, we derived a subset of the Corpus of Contemporary American English (CoCA; Davies, 2009). The CoCA consists of 485,179 texts published between 1990-2019. Texts are equally distributed across eight genres: academic, blog, fiction, magazines, newspapers, spoken, TV/movies, and web. We excluded the spoken and TV/movies genres to create a written English corpus of 416,401 documents (n = 641,410,953 words).

Language Measures

We measured lexical diversity using the measure of textual lexical diversity (MTLD; (McCarthy & Jarvis, 2010). MTLD is a robust measure of lexical diversity and, unlike type-token ratio (TTR), accounts for text length in calculating lexical diversity (McCarthy & Jarvis, 2010). To measure information density, we used Shannon entropy (K.T. Cunningham & Haley, 2020; Shannon & Weaver, 1949) as implemented in the ‘qdap’ package of the R programming language (Rinker, 2020). Shannon entropy is adapted from the field of information theory and asserts that in a given context, more unpredictable (i.e., less frequent) signals convey greater information than more predictable (i.e., more frequent) signals (Shannon & Weaver, 1949). Shannon entropy is calculated using the equation (Shannon & Weaver, 1949):

$$H(X) = -\sum_{i=1}^{n} P(X_i)\log_2 P(X_i)$$

Here, ‘information’ refers to the probability *P* of encountering a certain word X_i in a pool of *n* words. Essentially, the less probable the presence of the word, the more information it is judged to contain. Shannon entropy has a history of use in language research (Shannon & Weaver, 1949) and is sensitive to language processing in people with aphasia (K.T. Cunningham & Haley, 2020). In addition to our primary outcome variables of lexical diversity (McCarthy & Jarvis, 2010) and information density (Shannon & Weaver, 1964), we calculated the following variables, reporting overall means by document and means grouped by part of speech within document:

a) word count;

b) word length in number of letters;

c) syllable count;

d) syllables per word;

e) word age of acquisition (Kuperman et al., 2012);

f) word frequency (Brysbaert & New, 2009);

g) word concreteness (Brysbaert et al., 2014);

h) phonemes per word (Taylor et al., 2020).

We opted to include these psycholinguistic measures because a substantial body of evidence indicates that human language processing is sensitive to changes in these variables. For example, in a lexical decision task, reaction time is decreased for more frequent words and increased for less frequent words (Monsell et al., 1989). Similarly, people tend to process more concrete words (e.g., those representing concepts perceptible in the physical word) faster than less concrete and thus more abstract words such as *justice* (Brysbaert et al., 2014; Pexman et al., 2017). Such behavioral and physiological responses may be interpreted as increased cognitive effort in lexical processing secondary to specific distributions of psycholinguistic features characterizing a given word (Papesh & Goldinger, 2012; Piquado et al., 2010; van der Wel & van Steenbergen, 2018). Additionally, prior research suggests that, on average, written language tends to be characterized by a more 'lexically sophisticated' vocabulary than spoken language (e.g., lower word frequency, later age of acquisition). As such, we anticipated that this information would prove informative in developing the support vector machine learning algorithm we subsequently applied in our analysis.

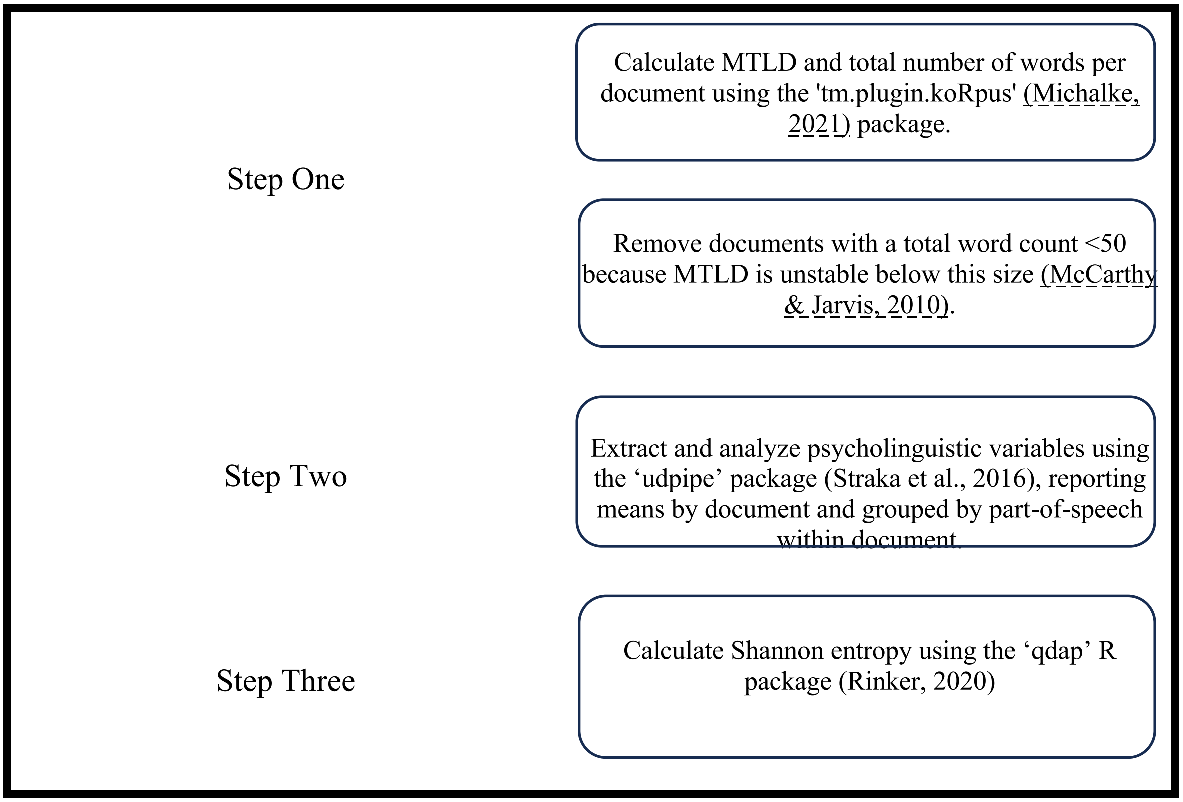
Data Processing

All data processing and analysis was completed using the Temple University High-Performance Computing Center 'compute' servers and/or the R programming language. The 'compute' server collection hosts 88 CPU cores and up to 1.5 TB of RAM across three interconnected remote servers.[1] After uploading our selected corpora to the server, we applied a custom cleaning pipeline (Finley, in revisions) to remove extraneous text (e.g., annotations) and non-alphabetic characters from all spoken and written

[1] See https://www.hpc.temple.edu/compute/ for details on 'compute' hardware and software features.

documents. Pre-processing steps varied between corpora secondary to differences in file types and structure (see Appendix A); however, the resultant cleaned dataset was homogenous across language modalities. Next, for all documents represented in each the spoken and the written corpus, we calculated our variables of interest using a set of custom natural language processing (NLP) pipelines. Figure 1 displays an overview of this process, separating components into consecutive steps.

Figure 1. Flowchart of processing steps used in extracting language measures from corpora


Step One
Calculate MTLD and total number of words per document using the 'tm.plugin.koRpus' (Michalke, 2021) package.
Remove documents with a total word count <50 because MTLD is unstable below this size (McCarthy & Jarvis, 2010).
Step Two
Extract and analyze psycholinguistic variables using the ‘udpipe’ package (Straka et al., 2016), reporting means by document and grouped by part-of-speech within document.
Step Three
Calculate Shannon entropy using the ‘qdap’ R package (Rinker, 2020)

Note. MTLD = Measure of Textual Lexical Diversity.

We concatenated results from our NLP analyses to create a dataset representative of each modality (e.g., spoken vs. written).

Analysis

To assess whether our measured language variables could be used to classify texts by modality, we used the 'e1071' R package to train a support vector machine classifier on 80% of the extracted language measures, with 20% reserved for testing (Meyer, 2023). In constructing a SVM classifier, it is critical to adjust model parameters in order to optimize performance. Prior to constructing the training and testing datasets, we sampled ~5% of the documents from each corpus and used ten-fold cross validation with varied model parameters to determine the optimal configuration for our spoken/written language classifier (Lantz, 2013; Meyer, 2023; Sirts et al., 2017; Yancheva & Rudzicz, 2016). Based on results from model tuning, we implemented a radial kernel SVM classifier with *Cost* = 100 and *gamma* = 0.1.

Results

Features of Spoken and Written English

We used the R programming language to analyze discourse measures and psycholinguistic features across a broad range of spoken and written English language contexts and genres, with the goal of determining whether differences in these variables could be used to accurately predict text modality. Table 3 provides an overview of the discourse measures and psycholinguistic features derived from each corpus after data preprocessing and cleaning. Visual inspection reveals that the testing and training sets, representing 80% and 20% of the overall data, appear to accurately reflect the distributional properties of the overall dataset along all measured variables.

Table 3. Overview of spoken and written language data

	Testing Set		Training Set		Overall	
	Spoken (n=19,229)	**Written (n=25,010)**	**Spoken (n=76,917)**	**Written (n=100,042)**	**Spoken (n=96,146)**	**Written (n=125,052)**
Word Count						
Mean (SD)	5930 (4350)	4490 (8240)	5970 (4340)	4530 (8480)	5960 (4340)	4520 (8430)
Median [Min, Max]	5360 [50.0, 29100]	2150 [50.0, 213000]	5410 [51.0, 45200]	2160 [50.0, 260000]	5400 [50.0, 45200]	2160 [50.0, 260000]
Word Length in Number of Letters						
Mean (SD)	3.82 (0.250)	4.44 (0.362)	3.82 (0.250)	4.45 (0.362)	3.82 (0.250)	4.44 (0.362)
Median [Min, Max]	3.75 [2.90, 5.45]	4.43 [2.67, 6.86]	3.75 [2.93, 5.74]	4.43 [2.99, 7.06]	3.75 [2.90, 5.74]	4.43 [2.67, 7.06]
Syllable Count						
Mean (SD)	7580 (5440)	6720 (11900)	7630 (5440)	6770 (12200)	7620 (5440)	6760 (12100)
Median [Min, Max]	6930 [62.0, 35100]	3240 [63.0, 279000]	6990 [54.0, 56800]	3260 [62.0, 346000]	6980 [54.0, 56800]	3260 [62.0, 346000]
Syllables per Word						
Mean (SD)	1.30 (0.0851)	1.50 (0.136)	1.30 (0.0853)	1.50 (0.136)	1.30 (0.0852)	1.50 (0.136)
Median [Min, Max]	1.28 [1.00, 1.92]	1.49 [1.05, 2.30]	1.28 [1.00, 2.03]	1.49 [1.02, 2.33]	1.28 [1.00, 2.03]	1.49 [1.02, 2.33]
Word Age of Acquisition						
Mean (SD)	5.05 (0.274)	5.62 (0.421)	5.05 (0.271)	5.62 (0.422)	5.05 (0.272)	5.62 (0.421)
Median [Min, Max]	4.98 [4.20, 7.24]	5.60 [4.29, 8.29]	4.98 [3.76, 7.46]	5.60 [4.16, 9.16]	4.98 [3.76, 7.46]	5.60 [4.16, 9.16]

	Testing Set		Training Set		Overall	
	Spoken (n=19,229)	**Written (n=25,010)**	**Spoken (n=76,917)**	**Written (n=100,042)**	**Spoken (n=96,146)**	**Written (n=125,052)**
Word Frequency						
Mean (SD)	7360 (665)	6110 (695)	7360 (665)	6120 (684)	7360 (665)	6120 (686)
Median [Min, Max]	7390 [918, 11600]	6140 [236, 14600]	7390 [2200, 13500]	6150 [97.8, 14000]	7390 [918, 13500]	6150 [97.8, 14600]
Word Concreteness						
Mean (SD)	2.50 (0.0947)	2.61 (0.150)	2.50 (0.0949)	2.61 (0.150)	2.50 (0.0949)	2.61 (0.150)
Median [Min, Max]	2.49 [2.12, 3.99]	2.59 [2.03, 4.14]	2.49 [2.16, 3.62]	2.59 [1.98, 4.40]	2.49 [2.12, 3.99]	2.59 [1.98, 4.40]
Phonemes per Word						
Mean (SD)	3.18 (0.228)	3.76 (0.341)	3.18 (0.228)	3.76 (0.342)	3.18 (0.228)	3.76 (0.342)
Median [Min, Max]	3.12 [2.28, 4.70]	3.75 [2.48, 5.47]	3.12 [2.31, 5.22]	3.75 [1.98, 5.70]	3.12 [2.28, 5.22]	3.75 [1.98, 5.70]
MTLD						
Mean (SD)	58.6 (20.6)	99.9 (26.3)	58.7 (20.9)	99.6 (27.0)	58.7 (20.9)	99.7 (26.8)
Median [Min, Max]	53.7 [3.90, 592]	98.3 [2.08, 463]	53.6 [3.03, 513]	97.8 [3.92, 1020]	53.6 [3.03, 592]	97.9 [2.08, 1020]
Shannon entropy						
Mean (SD)	5.42 (0.354)	5.74 (0.521)	5.43 (0.348)	5.74 (0.519)	5.43 (0.350)	5.74 (0.519)
Median [Min, Max]	5.49 [1.88, 6.60]	5.76 [1.99, 7.24]	5.49 [1.72, 6.70]	5.76 [2.10, 7.58]	5.49 [1.72, 6.70]	5.76 [1.99, 7.58]

Note. SD = standard deviation; Min = minimum; Max = maximum. Word count = mean word count per document; MTLD = Measure of Textual Lexical Diversity; Syllables = mean number of syllables per word; Letters = mean word length in letters; Concreteness =

	Testing Set		Training Set		Overall	
	Spoken (n=19,229)	Written (n=25,010)	Spoken (n=76,917)	Written (n=100,042)	Spoken (n=96,146)	Written (n=125,052)

mean per word, (Brysbaert et al., 2014); Age of Acquisition = mean per word, (Kuperman et al., 2012); Phonemes = mean per word; Frequency = word frequency, indexed as mean count per million words, from SUBTLEX-US (Brysbaert & New, 2009).

SVM Classification

We executed SVM training using ten-fold cross validation with model parameters set to *Cost* = 100 and *gamma* = 0.1. Classification accuracy for cross validations on the training dataset ranged from 92.86% - 93.45% (mean = 93.15%). Similar results were observed in model testing, where classification reached 93.15% accuracy (95% CI = [0.9291, 0.9338]). Table 4 shows model performance on the testing dataset.

Table 4. SVM classifier performance on testing set

		Actual		
		Spoken	**Written**	**Row Total**
Predicted	*Spoken*	17,185	988	18,173
	Written	2044	24033	26,066
Column Total		19,229	25,010	44,239

Table 5. SVM hyperplane parameters

Model input feature	Hyperplane parameter
Shannon entropy	-255.6736
MTLD	30.98
Word Count	-236.01
Mean Word Length	-554.2156
Syllables per Document	-268.3249
Syllables per Word	-29.55
Concreteness	190.24
Age of Acquisition	-120.72
Word Frequency	98.55
Phonemes per Word	-231.54
Number of letters per Word	-196.02

Note. MTLD = Measure of Textual Lexical Diversity.

Table 5 shows the hyperplane parameters for each of the SVM classifier input features. Hyperplane parameters are an abstracted representation of the relative contribution of each of the input features to the overall model performance. The greater the absolute value of the hyperplane parameter estimate, the greater its contribution to model performance. To determine the relative effects of each input variable on classification accuracy, we conducted a series of 'leave-one-out' analyses to assess the relative contribution of each of the model input features to the overall performance of the SVM. Across all ‘leave-one-out’ analyses, classification accuracy was consistently at or above 93%. This suggests that no single model input feature was driving classification performance. Rather, the combination of all input features synthesized to form an accurate prediction model.

Discussion

Prior research in language science has documented a relationship between WM and processes of spoken and written language production. WM in turn has been linked to two discourse constructs (i.e., idea density and lexical diversity) observed to vary within similar spoken and written language contexts (e.g., a picture description task). It is well-established that discourse measures and other language features (e.g., psycholinguistic variables) are influenced by communication contexts (e.g., in a classroom vs. a theme park). Advances in technology have enabled researchers to collect large and representative samples of naturalistic spoken language used across a wide range of contexts and genres, thereby allowing language contrasts by modality rather than situated within a given genre (e.g., books, magazines). In this project, we used natural language processing methods to extract salient discourse measures and psycholinguistic features from two large corpora representing spoken and written English, respectively. We aimed to determine whether written language, as suggested in prior research (Basso et al., 1978; Biber, 2004; Chafe & Tannen, 1987; Fergadiotis & Wright, 2011; Mitzner & Kemper, 2003), is characterized by greater information density and lexical diversity compared to spoken language. Using a SVM classification algorithm trained on a large number of discourse and psycholinguistic measures extracted from spoken (n = 76,917) and written (n=100,043) texts, we achieved ~93% accuracy in a binary classification task executed across 44,239 texts (n=25,010 written). To follow, we interpret these results, highlight study limitations, and offer directions for future research.

Interpretations

In this investigation, we collapsed across contexts and genres to examine effects of modality on language features previously linked to aspects of human language processing in both speaking and writing. The high degree of accuracy observed across multiple testing iterations of our classification algorithm suggests that modality contributes unique variance to the measured language features, independent of communicative context and/or genre. Considered within the broader context of research into spoken vs. written language, our results may be interpreted as an initial foundation of support for the previously proposed idea that cognitive systems (i.e., working memory) are differentially recruited in each modality, leading to downstream effects quantifiable in language features. However, results must be interpreted with caution given several study limitations that we address in the following section.

Limitations

Study findings suggest that spoken and written language are distinguishable based on language features extracted from large English-language corpora. We took great care in corpus selection to most accurately capture naturalistic use of spoken language across a range of communicative contexts. However, the fact remains that our spoken language measures are derived from podcasts, which may include pre-scripted elements (i.e., planning as in written language) or involve post-production editing, thereby distorting our language measures and calling into question their validity as representative of naturalistic spoken language (Clifton et al., 2020). We attempted to address this issue by comparing the Spotify Podcast Dataset with existent spoken language corpora (e.g., telephone calls, personal conversations) and found a favorable degree of homogeneity in dispersion of language measures across the comparison corpora. Given these findings, we argue that the language measures derived in our analyses are a valid representation of naturalistic spoken language.

Another limitation of the current study is the fact that metadata on linguistic subgenres (e.g., news, blogs) was available only for the written corpus; as such, we were unable to estimate interaction effects of modality*genre at a distributional scale. Our results indicate that modality does contribute some unique variance to language as measured by discourse features. However, it remains unclear whether effects of modality are further mediated by genre or other situational variables such as number of conversation partners. Finally, in this corpus analysis we were unable to determine whether idea density and lexical diversity are predicted by WM at the level of the individual. It is possible that WM is differentially enacted in producing spoken vs. written texts; however, cultivating a clearer understanding of the link between WM capacity, discourse features, and modality in a controlled setting (e.g., a laboratory experiment) is a critical next step in better understanding how cognition drives language at the level of the individual. We discuss this and other future directions in the following section.

Future Directions

In this project, we applied a binary SVM classification algorithm to evaluate distributional patterns of language measures in spoken vs. written English, achieving

~93% accuracy across model validation and testing. Our findings converge with prior research indicating that machine learning algorithms are capable of detecting even subtle changes in high-dimensional data, and as such have shown promise in accurately detecting disease from language data in schizophrenia and dementia, as well as predicting students' future academic performance (Crossley, 2020; Crossley & Allen, 2016; Elvevåg et al., 2007; Mitzner & Kemper, 2003; Mota et al., 2017; Paulsen et al., 1996). In analyzing features of spoken vs. written discourse, it is helpful to have a benchmark of expected performance: for example, procedural narrative elicitation (e.g., "Tell me how to tie a shoe,") generally yields language samples that are shorter as indexed by total number of words, elicit a less sophisticated vocabulary, and are less syntactically complex relative to, for example, a story retelling task (Fergadiotis & Wright, 2011; Stark, 2019). Performance outside of the expected range along a given discourse measure may signal underlying cognitive impairment or disease (Elvevåg et al., 2007; Fraser et al., 2015, 2019; Mitzner & Kemper, 2003; Mota et al., 2017; Orimaye et al., 2017; Sirts et al., 2017; Yancheva & Rudzicz, 2016). Thus, when indexing individual performance via language measures drawn from discourse, it is critical to interpret findings relative to the communicative context at hand. However, it remains unclear whether: a) modality differentially impacts lexical diversity and information density at the level of the individual; b) lexical diversity and information density are indeed driven by working memory, as previously proposed but (thanks to technological constraints) not yet empirically assessed across language modalities and genres. We begin investigating the link between WM, discourse features, and language modality in Experiment Two.

CHAPTER 3

INTERIM DISCUSSION

We will use the benchmarks described in this paper in Experiment 2 to contextualize discourse measures drawn from a behavioral study examining healthy undergraduates recruited from a large urban campus environment. Two of these indices (i.e., the Measure of Textual Lexical Diversity [MTLD], indexing lexical diversity; and information density, measured by Shannon entropy) are of particular interest vis-à-vis a transdisciplinary historical association with working memory (Chafe & Tannen, 1987; Mitzner & Kemper, 2003).

CHAPTER 4

EXPERIMENT 2

Abstract

Purpose: We predicted that working memory positively predicts information density and lexical diversity in narrative language samples; with stronger effects in spoken than in written language. **Method:** We elicited spoken and written language samples from healthy young adults after administering a comprehensive neuropsychological battery indexing working memory, vocabulary knowledge, and processing speed. Across two study visits occurring ~2-4 weeks apart, participants provided a spoken and a written response to each of four discourse elicitation prompts. We examined effects of modality on discourse features for each of two prompt categories (e.g., expositional and storytelling) with a particular focus on two discourse measures linked to WM: the Measure of Textual Lexical Diversity (MTLD) and Shannon entropy. **Results:** We found that MTLD but not Shannon entropy tends to vary with language modality. **Conclusions:** Although MTLD and Shannon entropy are highly correlated and both served as indices of *information density*, it seems that Shannon entropy does not index a similarly rich representation compared to MTLD in regards to the scope and depth of related representations in lexical, semantic, and phonological systems. Future studies will benefit from more ecologically valid discourse tasks and added measures (e.g., *theory of mind*).

Working Memory and Information Density in Speaking vs. Writing

Introduction

Corpus analysis and other approaches to computational linguistics serve to provide a broad overview of the attributes typifying a given language system. However, an understanding of the general patterns and features of a given language is insufficient to fully characterize language use at the individual level (Lim et al., 2020; Pexman & Yap, 2018). It is well established that a range of factors influence discourse measures (e.g., communicative context, modality). Individual level variance has historically been overlooked in neurotypical populations. Instead, neurotypical control groups are largely characterized *en masse* rather than seen for their potential to explain some of the variability surrounding influences on language production and downstream effects on discourse measures. In the current project, we explored the relationship between semantic processing and individual differences in working memory, processing speed, and vocabulary knowledge (i.e., semantic memory) in a cohort of healthy young adults enrolled in undergraduate coursework at a large public, urban university. Our goal in this investigation was to examine the relationship between cognition and language measures extracted from a discourse task. We draw upon cross-disciplinary perspectives to propose an integrated model of language processing in speaking and in writing, with a focus on the role of working memory in supporting language production in oral vs. written expression. Before presenting our model, we outline several cognitive systems thought to support language processing and well-known to vary at the level of the individual.

Language Processing in Speaking and Writing

Across theories of written and spoken language production, three cognitive processes are commonly identified: working memory, semantic memory (i.e., vocabulary size), and general cognitive ability (i.e., processing speed). To follow, we consider how current evidence supports or calls into question the structure and dynamics of these processes as related to language production. We then draw upon these findings to propose a simple model of language production, integrating oral and written modalities with a focus on the role of working memory in narrative language production.

Working Memory

Broadly, working memory refers to a process of item-specific activation and maintenance in order to complete some cognitive processing task. Over the years, different models of working memory (WM) have been proposed (Baddeley & Logie, 1999; Baddeley & Hitch, 1974; Cowan, 1999; Engle, 2002; R. C. Martin et al., 2020) and subsequently incorporated into models of spoken and written language production (Dell & Anderson, 2015; Flower & Hayes, 1981; N. Martin et al., 2018; R. C. Martin et al., 2020; Olive & Kellogg, 2002). Across models of working memory, there is general agreement that executive functions (e.g., attention, control) contribute to the selective activation and processing of items held online (Baddeley, 2003; Cowan, 2008; Engle, 2002). There is an ongoing debate regarding the structure and domain-specificity of the working memory system. For example, Baddeley and Hitch (1974) proposed a modular working memory system in which a domain-general "central executive" regulates the function of domain-specific subsystems (i.e., buffers). These buffers (e.g., visuospatial, phonological, and episodic) serve as temporary stores for items retrieved from long term memory (Baddeley, 2003; Baddeley & Logie, 1999). In contrast, Cowan's embedded processes model situates working memory within the long-term memory store (Cowan, 1988, 1998, 1999). A combination of voluntary and involuntary attentional processes serve to control item-specific selection from an activated portion of long-term memory, thereby bringing selected items into working memory for processing and manipulation (Cowan, 1999, 2008). While domain-specific portions of long-term memory may be activated, the embedded processes model characterizes working memory as a domain-general system. For the purposes of the present paper, we define working memory as the cognitive process by which items are activated and maintained online for processes of encoding and manipulation. This definition assumes overlap between working memory and short-term memory (STM) consistent with the structure proposed by Cowan (2008), in which WM is distinguished from STM by its use of higher-order executive functions to process and manipulate items held online. We used a set of three complex span tasks to assess working memory across multiple domains: Reading (e.g., verbal WM), Operations (e.g., non-verbal WM), and Symmetry, (e.g., visuospatial/non-verbal WM).

Working Memory in Oral vs. Written Expression

Successful language production requires integration across multiple sources of information (e.g., phonology, semantics, syntax) to complete activation and selection of a concept within a given communicative context. Once selected, motor pathways must be activated in order to speak or write (i.e., transcribe) the word associated with the chosen concept. Both oral and written language production models attribute this process to working memory (Dell & Anderson, 2015; Flower & Hayes, 1981; Kellogg et al., 2016; Olive & Kellogg, 2002). The motor sequences supporting written language are thought to tax WM capacity to a greater extent than those supporting spoken expression (Kellogg et al., 2016; Olive & Kellogg, 2002). Writing (e.g., transcription) not only takes more time than speaking, taxing WM duration, but also requires explicit instruction and practice to achieve fluent production. Children who participated in interventions targeting handwriting skills demonstrated post-treatment improvement in measures of written language quality (Alves et al., 2016; Alves & Limpo, 2015; Jones & Christensen, 1999). Alleviating transcription demands by implementing speech-to-text technology in children with traumatic brain injury (TBI) similarly resulted in improved language measures on a storytelling task (Noakes et al., 2019). This pattern of results indicates support for the role of WM in written language production: as transcription skills improve, more WM resources can be devoted to planning and lexical-semantic processing, leading to improved text quality (Alves et al., 2016; Jones & Christensen, 1999; Kellogg et al., 2016; Olive & Kellogg, 2002). Indeed, people with aphasia (PWA) have shown some benefit in these measures when producing personal narratives by writing vs. speaking (Behrns et al., 2009). Researchers have attributed this finding to the durability of written messages obviating the need to hold previously communicated information in WM. However, it is unclear whether working memory is taxed to a relatively greater degree in spoken or written language, and how the demands enacted within each modality influence language features.

Semantic Memory: Structure and Organization

In spoken and written language, words serve as symbolic representations of conceptual knowledge stored in semantic memory. Theories of semantic memory may be broadly categorized according their account of how conceptual knowledge is acquired and organized. Embodied approaches emphasize the role of direct experience in acquiring concept knowledge. These models propose that concept knowledge is stored in domain-specific sensorimotor neural regions, each connected to a central hub dedicated exclusively to language processing (Barsalou, 2016; Hoffman et al., 2018) Critics of embodied semantics argue that this approach fails to account for abstract concepts (e.g., justice, love) that cannot be grounded in sensorimotor experience. Feature-based models account for abstract and concrete concepts by proposing a vector structure of semantic memory. Words are represented along an array of vectors representing sensorimotor (e.g., shape, texture) and affective (e.g., funny, gross) features (Binder et al., 2016; Troche, 2018). Below, we give a brief overview of *concreteness* and other psycholinguistic variables thought to characterize this feature-based model of semantic memory.

Language Features: Psycholinguistic Measures

Researchers theorize that concepts sharing many features exist in close proximity to one another. Evidence for this relationship comes from studies of single-word processing tasks (e.g., lexical decision). In general, people tend to respond more quickly to words that are more closely related (Hoffman, 2018; Mirman & Graziano, 2012). This is true of words sharing similar sensorimotor and phonological/orthographic features (e.g., *fuzzy, furry*), as well as other psycholinguistic variables observed to influence human language processing:

a. Word frequency– reflects the relative frequency of word occurrences; measured using the SUBTLEX-US corpus, containing word frequency data derived from TV and movie subtitles (Brysbaert & New, 2009);

b. Word concreteness – measures the extent to which a given word may be experienced through sensorimotor modalities; indexed through aggregated crowd-sourced Likert ratings (Brysbaert et al., 2014);

c. Word age of acquisition – represented by aggregated human ratings reporting the age at which a given word was first encountered (Kuperman et al., 2012).

We include these descriptive measures along with other relevant variables (e.g., word length, phonemes; (Taylor et al., 2020) thought to influence language processing in speaking and/or writing. Semantic memory is commonly indexed through measures of vocabulary knowledge (Dunn, 2018; Kaya et al., 2012; Uttl, 2002).

Vocabulary Knowledge in Speaking and Writing

Conceptual knowledge is recruited in both spoken and written expression. However, the timing and means by which typically developing, speaking children acquire and process the symbolic representations of concepts differs between speaking and writing. Spoken language acquisition occurs incidentally as children are exposed to direct and indirect speech input from parents, caregivers, and siblings (Kuhl, 2000). Through this exposure, children learn to associate certain combinations of sounds (i.e., words) with concepts (Kuhl, 2000; Leonard et al., 2007). After vocabulary size reaches a certain threshold, this meaning mapping can occur with as little as one exposure to a novel word (e.g., “fast mapping,” Carey & Bartlett, 1978). In contrast, writing is taught incrementally, beginning with single alphabetic characters and building to words, sentences, and paragraphs (Alamargot et al., 2010; Jones & Christensen, 1999).

Written language instruction typically begins in preschool, after oral communication skills are established. Existing knowledge of oral word forms (i.e., phonological awareness) may be leveraged to facilitate written language development (e.g., a teacher may tell a child to "sound out" a word to figure out how to spell it). Greater phonological awareness is positively correlated with children's development of reading skills (Ehri et al., 2001). In turn, reading proficiency positively predicts time spent reading (i.e., print exposure), which is positively associated with writing proficiency into adulthood (Acheson et al., 2008; A. E. Cunningham & Stanovich, 1991; Epting et al., 2013; Mol & Bus, 2011). Given the association between print exposure and written expression, in our study we administered the Author Recognition Test [ART] (Acheson et al., 2008; Mar & Rain, 2015; Stanovich & West, 1989) to all participants as an index of print exposure.

In sum, converging evidence demonstrates the importance of exposure to written texts in developing writing proficiency. While spoken language exposure is virtually inevitable across many social contexts, exposure to significant amounts of written language (e.g., books, stories) requires an intentional allocation of time and resources. As such, individuals vary in their exposure to spoken and written forms of language, leading to differences in their oral vs. written vocabularies. For example, it is possible that a child has heard a word but not encountered it on the written page, or vice versa. The opaque orthography inherent to English systems of oral and written expression (e.g., phonological and orthographic representations do not have a 1:1 mapping) makes it possible that assessing vocabulary knowledge in a single modality is not sufficient to fully capture the store of conceptual knowledge located in semantic memory. Thus, it is important to index vocabulary knowledge not only through expressive language tasks (e.g., the short version of the North American Adult Reading Test [NAART35] (Uttl, 2002)), but also through receptive language tasks that do not require knowledge of written word forms (e.g., the Peabody Picture Vocabulary Test [PPVT] (Dunn, 2018)).

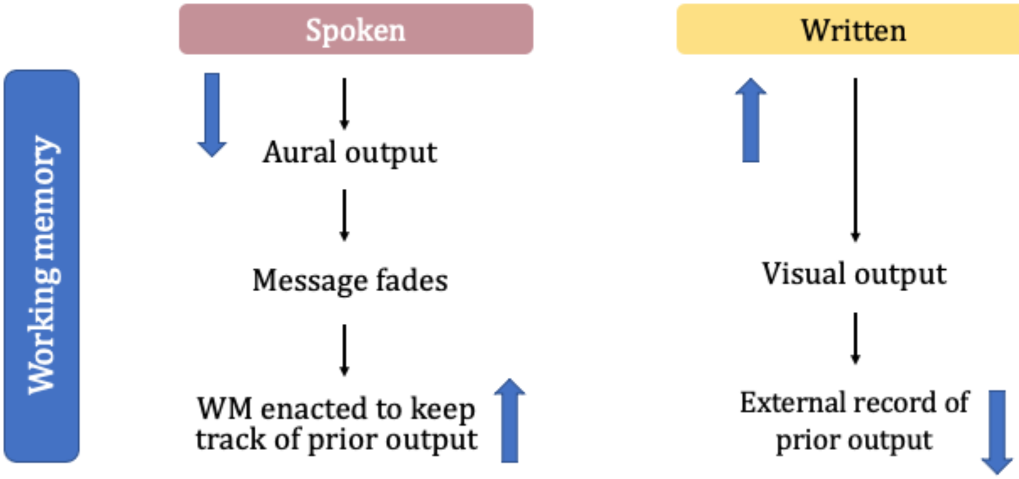
Processing Speed

Processing speed is most simply defined as the time it takes to complete some cognitive task. Together, processing speed and WM ability are seen as driving factors underlying measures of general fluid intelligence (Conway et al., 2002; Lee & Chabris, 2013). Faster processing speed is associated with better WM ability, a finding that has been attributed to similar demands of each enacted on processes of attention and inhibition (Conway et al., 2002; Lustig et al., 2006). It is well-established that processing speed varies according to task demands and stimulus characteristics (Brysbaert et al., 2000, 2018; Lustig et al., 2006; Monsell et al., 1989). In cognitive and language sciences, processing speed is often measured using response latency (e.g., reaction time) in single-word decision tasks (e.g., Balota et al., 2007; Pexman et al., 2017). Evidence shows that greater vocabulary knowledge corresponds with faster reaction time in single-word decision tasks (Pexman & Yap, 2018; Yap et al., 2012), although it is unknown what drives this behavioral observation (e.g., processing speed vs. WM). Processing speed is thought to support oral and written expression by facilitating processes of lexical-semantic activation and motor planning required in speaking and writing. In practice, it is unknown whether speaking or writing places greater demands on processing speed (e.g., relative to WM demands). However, it is likely that processing demands vary depending on communicative context. For example, a face-to-face spoken conversation may tax processing speed to a greater degree than writing a diary entry. To assess the role of processing speed and how it may relate to language features observed in spoken vs. written expression in the current study, we used the Connections Trail Making Tests (Salthouse, 2011).

Narrative Language Production: A Proposed Model Spanning Oral and Written Expression

In successful narrative expression, a variety of cognitive and linguistic systems must work in synchrony to retrieve lexical representations of semantic concepts, holding these items online until production can be completed, all while maintaining an overarching narrative theme or goal (Allen et al., 2016; Behrns et al., 2009; Chafe & Tannen, 1987; N. Martin et al., 2018). It is generally thought that in both oral and written expression, working memory supports this process of activation and maintenance as executive functions (e.g., cognitive control, attention) work to ensure appropriate selection and sequencing of words. However, differential demands are enacted on these systems in oral vs. written forms of expression (Behrns et al., 2009; Dell & Anderson, 2015; Kellogg, 2007; Olive & Kellogg, 2002). In spoken language, motor planning and execution of speech output is produced relatively quickly and once produced, is removed from focused attention, thereby freeing WM resources. In contrast, transcribing orthographic symbols via handwriting or typing occurs relatively slowly. Thus, WM demands for motor sequencing and execution are relatively lower in speaking and higher in writing (see Figure 2). This effect is attenuated by individuals' mastery of the transcription process – as transcription fluency improves, so too do measures indexing text quality (Alamargot et al., 2010; Alves et al., 2016; Alves & Limpo, 2015; Jones & Christensen, 1999; Olive & Kellogg, 2002).

Figure 2. Schematic overview of working memory demands in language production


Spoken
Written
Working memory
Aural output
Message fades
WM enacted to keep track of prior output
Visual output
External record of prior output

Note: Blue arrows represent the relative demands enacted on working memory through the language production process in spoken vs. written language. In motor planning and execution, working memory is taxed to a lesser extent in spoken relative to written language (upper blue arrows). However, after a message is produced, working memory demands are decreased in written relative to spoken language thanks to the durability of orthographic vs. aural representations (lower blue arrows).

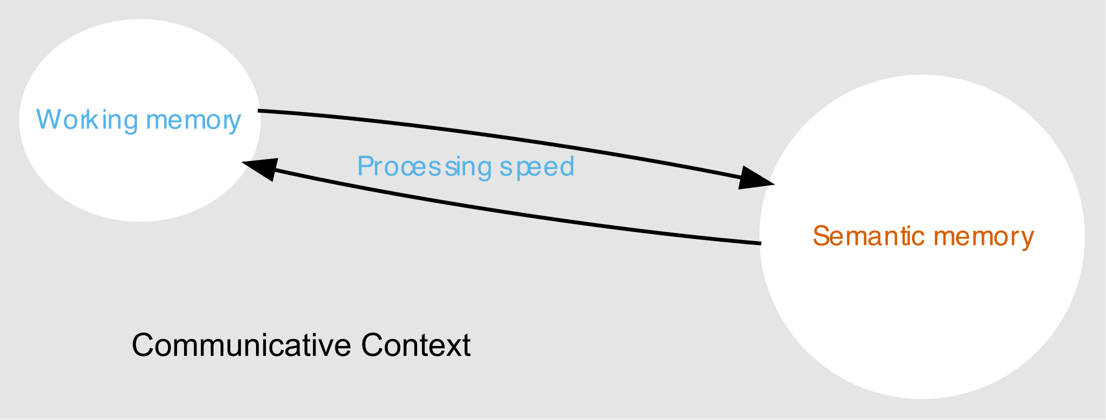
Pragmatic Influence

In typical spoken language contexts, there is a pragmatically induced time constraint on expression. Whether in a conversational or narrative (i.e., more monologic) context, extended pauses and halting language production are generally perceived unfavorably by audiences (Behrns et al., 2009) and may contribute to a speaker “losing the train of thought” secondary to overall slowing of the language production process (Dell & Anderson, 2015; Kellogg, 2007; Olive & Kellogg, 2002). In written language contexts, these acute temporal constraints are typically absent (Allen et al., 2016, 2019; Epting et al., 2013). Thus, although transcription of written language is on the whole slower than motor speech production, the lack of pragmatically-induced temporal pressures to respond leaves more opportunity to plan and revise throughout the written language production process (Epting et al., 2013; Flower & Hayes, 1981; Olive & Kellogg, 2002). The relative durability of written text relative to spoken text may further alleviate WM demands enacted in the language production process by maintaining an offline representation (e.g., a durable record) of language previously produced, rather than requiring online maintenance of aural output as in spoken language [see Figure 1] (Behrns et al., 2009; Epting et al., 2013). The opportunity for planning and revision in written expression may alter the interpretation of various discourse measures as indicators of cognitive-linguistic function when considered relative to interpretations that may be drawn from transcripts of oral language samples.

Summary and Overview of the Present Study

In general, written language tends to be characterized by greater *information density* than spoken language. *Information density* refers to the concentration of new information within a given text. It is possible that conveying information at a slower rate over the course of a spoken narrative (e.g., by using pauses or filler words) alleviates pragmatically-induced temporal demands on critical processes of selection and retrieval supported by working memory. *Information density* may be indexed in several ways; here, we use two overlapping but distinct indices: the Measure of Textual Lexical Diversity (MTLD) (McCarthy & Jarvis, 2010) and Shannon entropy (Shannon, 1950; Shannon & Weaver, 1949) to index *information density.* We propose that by repeatedly retrieving the same lexical item instead of activating and selecting a greater number of novel lexical items, it is possible that working memory demands for lexical retrieval are lessened in spoken relative to written English (see Figure 3). However, evidence to support a modality-mediated relationship between working memory and language is circumscribed to specific demographic groups (e.g., nuns) and discourse types (Fergadiotis & Wright, 2011; Mitzner & Kemper, 2003).

Figure 3. Graphical overview of key systems recruited in language production



Note: During language production, lexical-semantic items must be retrieved from semantic memory (i.e., lexical retrieval), a process mediated by general processing speed. Activated items are subsequently held online in working memory as an utterance unfolds, fading from working memory after successful production. Fluid cognitive processes are labeled in blue (e.g., working memory, processing speed) and crystalized aspects of cognition (e.g., semantic memory) are labeled in orange. How this process unfolds varies in different contexts (e.g., oral vs. written; formal vs. informal).

This experiment aims to elucidate the effect of modality (e.g., speaking vs writing) on working memory and language processing. We investigate the relationship between working memory and intra-individual narrative language measures in healthy young adults using a mix of expository and storytelling discourse elicitation tasks. We evaluate participants along a range of neuropsychological measures designed to probe the cognitive systems thought to contribute to language production in oral and written expression (i.e., working memory, vocabulary knowledge, processing speed). We examine the predictive relationship between participants' working memory score and two measures indexing *information density* (e.g., MTLD and Shannon entropy), using linear-mixed effects models and canonical correlation analysis to test our prediction that working memory positively predicts information density and lexical diversity in narrative language. However, we anticipate that this effect will be relatively greater in spoken vs. written language.

Methods

Participants

Of n=36 participants recruited to the study, n=5 participants were excluded from analysis secondary to incomplete data (e.g., attended only one study visit; n=2) or ineligibility based on inclusion criteria (n=3). An additional n=8 participants were enrolled into the study and at the time of writing this document, are currently going through the study protocol. A total of 23 healthy young adults (22 F); ranging in age from 18 - 32 years (M=21, SD = 2.61) participated in the study. Inclusion criteria were as follows: native English speaking with fluency in spoken and written communication, enrolled in university coursework at the undergraduate level, age between 18 – 35 years, normal or corrected-to-normal vision, normal or corrected-to-normal hearing. Exclusion criteria included the presence and/or history of any of the following: neurological disorder or injury (e.g., TBI, stroke, concussion), neurodegenerative disease, language or cognitive disorder, first language other than English. Participants were recruited via virtual flyers shared on undergraduate course websites and via email.

Materials

All participants were administered a neuropsychological battery designed to assess the following cognitive areas: working memory, vocabulary knowledge, processing speed, and print exposure. To follow is a brief overview of the neuropsychological assessments used to measure participants' abilities in each of these cognitive domains.

Working Memory

Working memory was assessed using a set of complex span tasks (Unsworth et al., 2005). In a complex span task, working memory is targeted by interspersing distractors (e.g., sentences, math problems) among target items in a serial recall task (Daneman & Carpenter, 1980). Task difficulty increases as the number of target items and distractors increases. Complex span tasks are used (Conway et al., 2005; Unsworth et al., 2005; Wilhelm et al., 2013) to assess working memory in a variety of domains (e.g., verbal, spatial). We used a computerized, automated version of the complex span task that demonstrates good reliability and validity across over 5,000 trials (Redick et al., 2012; Unsworth et al., 2005). Participants completed three complex span tasks, each targeting a different working memory domain: Reading, Operations, and Symmetry. We derived a Composite working memory score for each participant by first z-scaling then averaging across the domain-specific complex span task scores. Although we anticipated that the Reading Complex Span Score would be the best predictor of our outcome variables, we also anticipated that the domain-specific working memory scores would be highly correlated for each participant. Thus, we used correlational analyses to determine which working memory measure (i.e., Reading, Operations, Symmetry, or Composite) to use in linear mixed-effects modeling.

Vocabulary Knowledge

We measured vocabulary knowledge using the short version of the North American Adult Reading Test (NAART35; (Uttl, 2002). In NAART35 administration, vocabulary knowledge is assessed by having participants read aloud a list of 35 irregularly spelled English words (e.g., abstemious, demesne). One point is awarded for each correct pronunciation, with a total of 35 possible points. The NAART35 demonstrates good reliability (Cronbach's α = 0.93) and validity (correlation r = 0.76 with the Wechsler Adult Intelligence Scale – Revised Vocabulary raw score WAIS-R Vocabulary, (Wechsler, 1981)); and correlation r = 0.98 with the full NAART (Blair & Spreen, 1989) across young, middle-aged, and old adults (Uttl, 2002). The NAART35 has previously been used in research examining individual differences in language processing (Pexman & Yap, 2018).

We measured receptive vocabulary using the Peabody Picture Vocabulary Test - Fifth Edition (PPVT-5; Dunn, 2018). The PPVT-5 is a widely used picture naming assessment, normed on an English-speaking US population ranging in age from 2;6 - 99;0 (years; months). Mean split-half reliability for the PPVT-5 = .97, with a standard error of measurement (SOM) = 2.63. The PPVT-5 demonstrates good validity across a range of comparable objective measures of receptive and expressive language abilities, respectively (Dunn, 2018).

Processing Speed

To assess processing speed, all participants completed the Connections Trail-Making Tests versions A and B (Salthouse, 2011). In these timed tests, participants are required to complete alternating (i.e., Version A) or non-alternating (i.e., Version B) sequential trails connecting letters, numbers, or both. Structural equation modeling and contextual analyses indicate that performance on both versions of the test reflects general fluid cognition and processing speed (Salthouse, 2011). The alternating version of the test is primarily related to cognitive processing speed, with some influence of general fluid cognitive ability. The non-alternating version of the test accounts for additional variance in general cognitive ability. In the current study, cognitive processing speed was indexed as the difference in score between the alternating and the non-alternating trail-making tasks (Salthouse, 2011).

Print Exposure

We indexed print exposure using the Author Recognition Test (ART; Acheson et al., 2008; Mar & Rain, 2015; Stanovich & West, 1989). In this task, participants read a list of names and place a check mark next to the names they know to be writers. Participants are discouraged from guessing in this signal-detection paradigm, where a final score is calculated by subtracting the total number of incorrect answers from the total number of correct answers. The version of the ART used in the present study contains 160 author names and 40 foils (Mar & Rain, 2015).

Narrative Elicitation Prompts

We administered four different prompts to participants: two expositional prompts and two narrative prompts. Three of the four prompts were derived from the AphasiaBank protocol: the broken window (BW) story, the cat rescue story, and the Cinderella story (MacWhinney et al., 2011). The BW story and the cat rescue story represented the two expositional prompts, while the Cinderella story was one of the two narrative prompts. We selected these prompts because they are widely used among language researchers, offer a standard approach to language sample elicitation and evaluation, and their respective effects on text-based language measures are recently established within the AphasiaBank data set (MacWhinney et al., 2011; Stark, 2019). The second narrative prompt came from the short film *Snack Attack*. *Snack Attack* is a 4-minute, 35-second animated film, summarized by its creators: “Waiting to board the train, an old lady just wants to eat her cookies in peace, but hijinks ensue when a teenager on the platform next to her seems intent on sharing them too.”[2] We included a film-based stimulus because prior work has criticized Cinderella for being too simplistic and suggested use of more complex and ecologically valid discourse stimuli (K. T. Cunningham & Haley, 2020). Additionally, prior studies of information density in discourse have successfully used wordless short films in a similar manner to elicit narrative language from healthy controls (Ravid & Berman, 2006). The short film selected for the proposed project represents a simple but engaging story depicting familiar activities (e.g., buying a snack, waiting for public transit) likely to be familiar to all participants.

Experimental Procedures

Study participation involved two visits, spaced 2-4 weeks apart (mean = 21, SD = 4.26). In the first visit, all screening, consent, and cognitive assessment procedures were administered by a trained research assistant in a quiet room located in the Department of Communication Sciences and Disorders. Participants underwent screening for inclusion/exclusion criteria and completed a verbal informed consent procedure prior to beginning any study tasks. All complex span tasks were presented on a computer screen using EPrime 3.0 Professional software (Psychology Software Tools, Incorporated); all other testing was completed via pen-and-paper. All study procedures were audio recorded. Videorecording was used only during the storytelling portion of each visit. The

[2] *http://snackattackmovie.com/#about%20CastCrew*

second study visit included only the storytelling task activities. Participants were able to opt whether to complete the second study visit in-person or remotely via Zoom videoconferencing software.

Table 6. Story modality by prompt across task conditions

Condition		**Prompt Category**			
		Expository		**Story (re)telling**	
		Broken Window	*Cat rescue*	*Cinderella*	*Snack*
AA	*Visit 1*	Spoken	Written	Spoken	Written
	Visit 2	Written	Spoken	Written	Spoken
AB	*Visit 1*	Spoken	Written	Written	Spoken
	Visit 2	Written	Spoken	Spoken	Written
BA	*Visit 1*	Written	Spoken	Spoken	Written
	Visit 2	Spoken	Written	Written	Spoken
BB	*Visit 1*	Written	Spoken	Written	Spoken
	Visit 2	Spoken	Written	Spoken	Written

Note. ‘AA,’ ‘AB,’ ‘BA,’ and ‘BB’ are labels for task conditions.

Data Processing

Spoken Language Transcription. We used automated speech-to-text software (Otter Software Tools, 2023) to transcribe participants’ spoken language samples. To verify transcription accuracy, we compared all automatically-transcribed files with the recorded audiovisual file, correcting any perceived errors.

Narrative Data Preprocessing. Using a custom text processing pipeline in the ‘R’ programming language[3], we transformed participant narratives into ordered vectors of single words. Prior to conducting our analyses, we removed narratives that contained

[3] Hosted on the ‘compute’ high-performance computing server group along with the other technological resources described in Experiment 1.

>50% off topic references. To determine off topic references, we used NLP methods to generate an index of story-specific semantic information content. Researchers (N. Martin et al., 2020; Richardson et al., 2021) have measured story-specific information by manually identifying and counting the number of “correct information units” (CIUs; Nicholas & Brookshire, 1993) or “main concepts” (Kong et al., 2016) present in a narrative. Here, we used the R programming language to automatically identify information content units characteristic of spoken and written participant narratives. Our approach is modeled on prior discourse analysis work (Sirts et al., 2017; Yancheva & Rudzicz, 2016) examining biomarkers of dementia using data from DementiaBank (Lanzi et al., 2023).

First, we matched each word to its associated word embedding indexed in the Wikipedia 2014 GloVe dataset (Pennington et al., 2014). In word embedding models, meaning is mathematically modeled as a hyperparameter across a number of abstract dimensions (n=50 in GloVe). GloVe represents word meaning based on word co-occurrence statistics within a user-specified window size. This may be thought of as a more complex version of relatively simpler co-occurrence based representations of word meaning (e.g., latent semantic analysis; (Landauer & Dumais, 1997; Pennington et al., 2014). Next, we filtered the word embedding data to include only nouns and verbs (Sirts et al., 2017; Yancheva & Rudzicz, 2016). After matching each word to its associated word embedding in the Wikipedia 2014 GloVe dataset (Pennington et al., 2014), we used the ‘stats’ package to conduct a k means cluster analysis, determining optimal cluster size for each combination of prompt*modality using the elbow method. Across conditions, results suggested an optimal cluster size *k*=10. We repeated this analysis using a ‘leave-one-out' approach, iterating across participants to determine the percentage of off-topic references produced by any one participant relative to the clusters generated by the *n*-1 sample. Participants with >50% word embeddings falling outside of the generated ICU clusters were excluded from analysis.

Outcome Measures

Our primary outcome measures were information density and lexical diversity. We measured information density using Shannon entropy (Ravid & Berman, 2006; Shannon & Weaver, 1949) as implemented in the 'qdap' package of the R programming language. Shannon entropy is a linguistic measure originated in the field of information theory and asserts that in a given context, more unfamiliar signals (i.e., words) convey greater information than more familiar signals (Shannon & Weaver, 1949). Similar to propositional density, Shannon entropy does not capture story-specific semantic content. Rather, Shannon entropy indexes informativity as the relative degree of predictability of a given word form (e.g., the probability of encountering word X_i in context n (Shannon & Weaver, 1949).

To measure lexical diversity, we used the 'tm.plugin.koRpus' package (Michalke, 2021) to derive the measure of textual lexical diversity (MTLD) for each narrative (McCarthy & Jarvis, 2010). MTLD is adapted from the type-token ratio (TTR; e.g., the ratio of unique words to total words in a given language sample). However, unlike TTR, it is robust to effects of text length. This is achieved by using a flexibly-sized moving window to calculate TTR across a language sample until it achieves a threshold of 0.72 (McCarthy, 2005; McCarthy & Jarvis, 2010), at which point the number of words in the window is saved and the process is repeated (Figure 4). Each iteration yields an additional factor score. In the event that a given language sample *LS* ends with a word string such that TTR$_{CURRENT}$ > TTR$_{CRITERION,}$ a partial factor score *PF* is calculated based on the percent change needed such that TTR$_{CURRENT}$ = TTR$_{CRITERION:}$

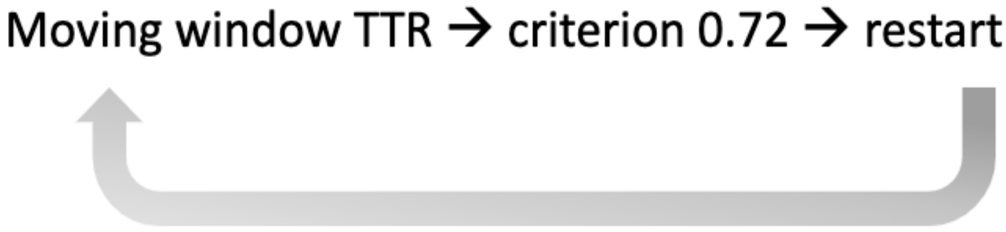
$$PF = \frac{\text{TTR\_CURRENT}}{1-\text{TTR\_CRITERION}}$$

MTLD is calculated by dividing the total number of words in a given language sample (TNW_{LS}) by the sum of all whole factors + *PF* (i.e., the total factor score, *FACTOR_TOTAL)* such that:

$$MTLD_{LS} = \frac{\text{TNW\_LS}}{\text{FACTOR\_TOTAL}}$$

In other words, for any given language sample of length *N* (measured in total number of words), the greater the number of words required to gain an additional factor score (i.e., achieve the criterion TTR), the greater the lexical diversity and the lower the number of factors yielded in analysis. Broadly, MTLD represents lexical diversity as the ratio of total number of words to the number of factors calculated from a set of *N* words (McCarthy, 2005).

Figure 4. The measure of textual lexical diversity – factor scoring



Note. TTR = Type-Token Ratio.

Psycholinguistic Features

In addition to our primary outcome measures, we calculated the following psycholinguistic variables for each language sample:

a. word frequency (Brysbaert & New, 2009);
b. word concreteness (Brysbaert et al., 2014);
c. word length in number of letters;
d. number of phonemes per word (Taylor et al., 2020);
e. word age of acquisition (Kuperman et al., 2012);
f. number of phonemes per document (Taylor et al., 2020).

We included these language measures because they are widely used in language research and abundant evidence suggests these features influence human language processing. In our analysis, we examined participant performance along these measures within context of the associated benchmarking estimates derived for spoken and written English in Experiment 1.

Table 7. Proposed linear mixed-effects models

Predictor Variable	Equation
MTLD	*Information density ~ working memory + modality + working memory*modality + (1|participant) + (1 | prompt_category) + (1|prompt) + error*
Shannon entropy	*Lexical diversity ~ working memory + modality + working memory*modality + (1|participant) + (1 | prompt_category) + (1|prompt) + error*

Note. MTLD = Measure of Textural Lexical Diversity.

Statistical Analysis

We used linear mixed-effects modeling to test our primary prediction that working memory would positively predict lexical diversity and information density with stronger effects in written than spoken language. Linear mixed-effects models (LMMs) allow for consideration of both random and fixed effects, making them powerful tools to assess intra-individual performance across repeated measures (Magezi, 2015; Wiley & Rapp, 2019; Winter, 2013). We ran two LMMs to predict information density and lexical diversity (see Table 7 for model overviews). In both models, we included working memory as a predictor variable, along with modality and an interaction effect of working memory*modality. We included participant, prompt, and prompt category as random effects. We ran several LMMs with additional predictors of interest (e.g., vocabulary knowledge, processing speed, print exposure) added in a stepwise fashion with the goal of optimizing variance accounted for while avoiding multicollinearity among model inputs. Among these, we maintained variables that were significant predictors at $p \leq 0.05$.

Canonical Correlation Analysis. In addition to addressing our primary prediction that working memory has a greater effect on information density in spoken vs. written language, we also investigated whether the previously-reported relationship between working memory and our primary outcome variables (e.g., information density and lexical diversity) was similarly mediated by language modality across our spoken and written language sample distributions. Canonical correlation analysis (CCA) is a widely-used statistical approach to dimension reduction and represents a multivariate generalization of the Pearson product moment correlation (Altaf et al., 2020; Dattalo, 2014; Ho, 1987; Iweka & Anthonia, 2018; LeClere, 2006). CCA models the relationship between two datasets, each containing ≥ 2 variables, by assessing their respective variance and covariance matrices. The goal of CCA is to determine the linear combination of variables that maximizes the shared variance between the two datasets.

The number of *canonical correlations* yielded by this process is typically constrained by the number of variables in the smaller dataset (Altaf et al., 2020; Dattalo, 2014; Gonzalez et al., 2008; Iweka & Anthonia, 2018; LeClere, 2006). We constructed a predictor dataset consisting of our three working memory measures, along with an outcome dataset containing MTLD and Shannon entropy data, then used the 'stats' R package to complete canonical correlation analyses for participants' spoken and written language samples. For each predictor-DV pair, we compared the correlation coefficients from the canonical correlation analysis with beta weights produced by a simple linear regression model.

Results

Neuropsychological Battery

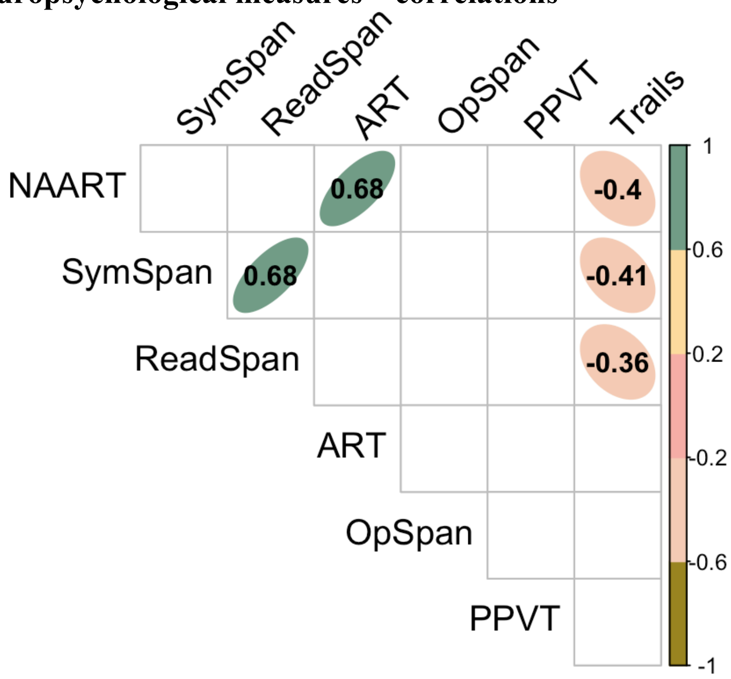
Summary statistics for neuropsychological test battery results are displayed in Table 8. Participant performance across the NAART35 (Muraki & Pexman, 2021), the ART (Stanovich & West, 1989), and the PPVT (Dunn, 2018) was within the expected range of performance based on normative values and/or previously published results for a similar demographic group. Performance on the complex span tasks (e.g., Symbol Span, Reading Span, Operation Span) was similarly within the range of expected performance based on normative data (Redick et al., 2012). Across the trail-making tasks (e.g., Trails A, Trails B, Trails B-A), participants in our study were on average somewhat slower in task completion relative to previously published results; however, the range of completion times across our sample was generally consistent with expected results (Tombaugh, 2004). Figure 5 displays bivariate correlations (significant at $p > 0.05)$ among the neuropsychological measures of interest to the current investigation. "Trails" refers to the Trails B-A difference score, used to index processing speed.

Table 8. Neuropsychological evaluation outcomes

	Standard Score (N=23)	z-score (N=23)
NAART		
Mean (SD)	16.0 (5.83)	0.0874 (0.980)
Median [Min, Max]	18.0 [3.00, 26.0]	0.423 [-2.10, 1.77]
ART		
Mean (SD)	9.22 (4.49)	0.0918 (0.988)
Median [Min, Max]	9.00 [2.00, 17.0]	0.0440 [-1.50, 1.80]
PPVT		
Mean (SD)	100 (10.1)	0.104 (0.975)
Median [Min, Max]	100 [83.0, 123]	0.0658 [-1.58, 2.29]
Symbol Span		
Mean (SD)	30.9 (6.22)	-0.0201 (0.995)
Median [Min, Max]	29.0 [17.0, 42.0]	-0.326 [-2.25, 1.76]
Reading Span		
Mean (SD)	59.0 (11.8)	0.187 (0.792)
Median [Min, Max]	60.0 [32.0, 75.0]	0.254 [-1.63, 1.26]
Operation Span		
Mean (SD)	59.7 (13.7)	0.147 (0.912)
Median [Min, Max]	62.0 [15.0, 75.0]	0.298 [-2.84, 1.16]
Trails A		
Mean (SD)	28.9 (10.7)	0.00862 (1.03)
Median [Min, Max]	25.8 [14.1, 57.8]	-0.288 [-1.41, 2.79]
Trails B		
Mean (SD)	64.4 (23.8)	-0.0130 (1.02)
Median [Min, Max]	62.1 [33.6, 135]	-0.110 [-1.33, 3.00]
Trails B-A		
Mean (SD)	35.5 (18.3)	-0.0219 (1.02)
Median [Min, Max]	31.6 [12.6, 88.2]	-0.239 [-1.30, 2.92]

Note. Min = Minimum; Max = Maximum; SD = Standard deviation. NAART = North American Adult Reading Test; ART = Author Recognition Test; PPVT = Peabody Picture Vocabulary Test-5; Symbol Span, Operation Span, and Reading Span refer to the complex span task subtype scores; Trails = Connections Trail Making Test.

Figure 5. Neuropsychological measures – correlations



Note. Only correlations significant at *p> 0.05* are displayed. NAART = North American Adult Reading Test; ART = Author Recognition Test; PPVT = Peabody Picture Vocabulary Test-5; Symbol Span, Operation Span, and Reading Span refer to the complex span task subtype scores; Trails = Connections Trail Making Test.

Narrative Data

After running participant narratives through our data cleaning pipelines in R, we calculated our variables of interest and aggregated results in several ways in order to tease out differential performance as an effect of modality and prompt category (Table 9). To follow, we briefly summarize the results presented in these tables before presenting the results of our linear mixed-effects model analyses.

Table 9. Discourse measures by modality and prompt category

	Expositional		**Storytelling**		**Overall**	
	Spoken (n=46)	**Written (n=46)**	**Spoken (n=46)**	**Written (n=46)**	**Spoken (n=92)**	**Written (=92)**
Word Count						

	Expositional		Storytelling		Overall	
	Spoken (n=46)	Written (n=46)	Spoken (n=46)	Written (n=46)	Spoken (n=92)	Written (=92)
Mean (SD)	246 (325)	130 (63.2)	625 (284)	310 (73.7)	436 (358)	220 (113)
Median [Min, Max]	161 [47.0, 1730]	114 [33.0, 286]	602 [209, 1710]	291 [163, 540]	344 [47.0, 1730]	236 [33.0, 540]
Word Length in Number of Letters						
Mean (SD)	3.75 (0.183)	3.97 (0.254)	3.91 (0.192)	4.14 (0.236)	3.83 (0.205)	4.05 (0.258)
Median [Min, Max]	3.73 [3.35, 4.13]	3.92 [3.44, 4.60]	3.89 [3.47, 4.31]	4.12 [3.67, 4.61]	3.83 [3.35, 4.31]	4.02 [3.44, 4.61]
Syllable Count						
Mean (SD)	295 (384)	161 (79.5)	784 (365)	405 (99.7)	540 (447)	283 (152)
Median [Min, Max]	194 [56.0, 2050]	141 [44.0, 375]	760 [247, 2170]	391 [213, 742]	406 [56.0, 2170]	295 [44.0, 742]
Syllables per Word						
Mean (SD)	1.20 (0.0481)	1.24 (0.0648)	1.25 (0.0574)	1.31 (0.0789)	1.22 (0.0583)	1.27 (0.0794)
Median [Min, Max]	1.19 [1.11, 1.32]	1.25 [1.12, 1.46]	1.25 [1.17, 1.38]	1.30 [1.17, 1.48]	1.22 [1.11, 1.38]	1.26 [1.12, 1.48]
Word Age of Acquisition						
Mean (SD)	4.39 (0.145)	4.42 (0.183)	4.62 (0.0905)	4.65 (0.129)	4.50 (0.166)	4.53 (0.196)
Median [Min, Max]	4.38 [4.06, 4.85]	4.40 [4.11, 4.82]	4.62 [4.45, 4.83]	4.63 [4.45, 5.10]	4.53 [4.06, 4.85]	4.56 [4.11, 5.10]
Word Frequency						
Mean (SD)	7570 (767)	7280 (1140)	6750 (456)	6430 (405)	7160 (752)	6850 (951)

	Expositional		Storytelling		Overall	
	Spoken (n=46)	**Written (n=46)**	**Spoken (n=46)**	**Written (n=46)**	**Spoken (n=92)**	**Written (=92)**
Median [Min, Max]	7570 [5720, 9510]	7280 [4860, 10700]	6640 [6000, 7680]	6360 [5430, 7530]	7090 [5720, 9510]	6630 [4860, 10700]
Word Concreteness						
Mean (SD)	2.79 (0.135)	2.90 (0.138)	2.62 (0.0716)	2.74 (0.0756)	2.70 (0.138)	2.82 (0.135)
Median [Min, Max]	2.77 [2.52, 3.13]	2.91 [2.54, 3.22]	2.62 [2.47, 2.79]	2.75 [2.51, 2.88]	2.68 [2.47, 3.13]	2.82 [2.51, 3.22]
Phonemes per Word						
Mean (SD)	2.93 (0.125)	3.06 (0.172)	3.03 (0.170)	3.19 (0.224)	2.98 (0.156)	3.12 (0.209)
Median [Min, Max]	2.92 [2.72, 3.42]	3.00 [2.78, 3.57]	3.01 [2.73, 3.36]	3.13 [2.83, 3.63]	2.94 [2.72, 3.42]	3.07 [2.78, 3.63]
MTLD						
Mean (SD)	40.2 (9.53)	48.2 (14.9)	42.1 (6.83)	53.7 (10.5)	41.1 (8.30)	51.0 (13.1)
Median [Min, Max]	41.4 [20.5, 66.7]	46.5 [21.0, 85.2]	41.9 [29.6, 60.4]	53.8 [34.3, 86.9]	41.6 [20.5, 66.7]	51.0 [21.0, 86.9]
Shannon entropy						
Mean (SD)	3.99 (0.386)	3.86 (0.397)	4.56 (0.219)	4.46 (0.153)	4.28 (0.423)	4.16 (0.422)
Median [Min, Max]	4.03 [2.95, 4.97]	3.89 [2.98, 4.56]	4.57 [4.16, 5.03]	4.47 [4.08, 4.77]	4.30 [2.95, 5.03]	4.31 [2.98, 4.77]

Note. MTLD = Measure of Textual Lexical Diversity. Min = Minimum; Max = Maximum; SD = Standard deviation.

Across both expositional and storytelling prompts, MTLD in written narratives tended to be greater than in spoken narratives ($MTLD_W > MTLD_S$). MTLD was greatest for written storytelling prompts and least for spoken expositional narratives. Prompt-based discrepancies in MTLD were greater in written narratives than in spoken narratives (*mean difference$_W$* = 5.5; *mean differences$_S$* = 1.9). Shannon entropy was greater in storytelling vs. expositional prompts, with negligible differences observed by modality within prompt category. Spoken narratives tended to be longer (e.g., Word Count; Syllable Count) than written narratives for both prompt categories. Expositional prompts were shorter than storytelling prompts, regardless of modality. Mean AoA was slightly greater for storytelling vs. expositional prompts and did not vary greatly by modality within prompt category. Mean word frequency was somewhat greater for expositional vs. storytelling narratives, with slightly reduced word frequency observed in written vs. spoken narratives for both expositional and storytelling prompts.

Examining overall effects of modality, we found that among spoken language samples, MTLD was on average less than in the written group (*mean difference* = 10.1), with a more restricted range of variability also observed (*MTLD range$_S$* = 46.2; *MTLD range$_W$* = 65.9). Turning to Shannon entropy, we see a different pattern of results. Table 4 shows that overall Shannon entropy was homogenous across modalities at the distributional level (*mean difference* = 0.12). In a *post hoc* analysis, we used paired samples t-tests to contrast participants' overall means by modality for MTLD and Shannon entropy. The general pattern of results were consistent with those presented in Table 4: significant values were observed only for MTLD (*mean difference* = 9.82, *t* = 5.73, *degrees of freedom (df)* = 22; *p <0.001),* while Shannon entropy did not vary significantly by modality *(mean difference* = -0.11; *t*= -4.1, *df*= 22; *p = 1*). Similar homogeneity between groups was observed across all psycholinguistic variables except those measuring text length (e.g., Word Count, Syllable Count). Indexed along these variables, spoken language samples were on average longer than written language samples and demonstrated greater variability. Along measures of Word Count and Syllable Count, Shannon entropy appears to be more variable in spoken than in written language. It is possible that the observed differences in Shannon entropy and MTLD are related to discrepancies in the use of “filler” words (e.g., *um, uh*) often encountered in spoken language contexts (Zhu et al., 2022) and thought to be used to offset working memory load putatively associated with spoken language processing (Biber, 1986; Chafe & Tannen, 1987). To explore this possibility, we repeated our measures after excluding the filler words *um, umm, uh, uhh* (Zhu et al., 2022). We observed little to no changes in our measured variables as a result of this manipulation (see Table 10), results reaffirmed in Welch’s two sample t-tests contrasting participant mean values measured with and without filler words (MTLD: $t_S = 0.30$; $df = 43.9$; $p = 0.76$; $t_W = -0.06$; $df = 43.9$; $p = 0.95$. Shannon entropy: : $t_S = 0.11$; $df = 43.9$; $p = 0.91$; $t_W = -0.05$; $df = 44$; $p = 0.96$).

Table 10. MTLD and Shannon entropy measured with and without filler words

	Spoken **(N=92)**	**Written** **(N=92)**
Word Count		
Mean (SD)	436 (358)	220 (113)
Median [Min, Max]	344 [47.0, 1730]	236 [33.0, 540]
Word Count_NoFill		
Mean (SD)	431 (357)	220 (114)
Median [Min, Max]	342 [46.0, 1710]	236 [33.0, 540]
MTLD		
Mean (SD)	41.1 (8.30)	51.0 (13.1)
Median [Min, Max]	41.6 [20.5, 66.7]	51.0 [21.0, 86.9]
MTLD_NoFill		

	Spoken (N=92)	Written (N=92)
Mean (SD)	40.6 (8.05)	51.1 (13.1)
Median [Min, Max]	41.1 [18.8, 66.4]	51.2 [21.0, 87.6]
Shannon entropy		
Mean (SD)	4.28 (0.423)	4.16 (0.422)
Median [Min, Max]	4.30 [2.95, 5.03]	4.31 [2.98, 4.77]
Shannon_NoFill		
Mean (SD)	4.27 (0.429)	4.16 (0.423)
Median [Min, Max]	4.29 [2.91, 5.03]	4.31 [2.98, 4.77]

Note. MTLD = Measure of Textual Lexical Diversity; NoFill = Calculated from language samples after removing *um, umm, uh, uhh.*

In this investigation, we were primarily concerned with the effects of modality on discourse measures independent of context or genre; however, it is well-established that prompt category influences language features in intra-individual narrative discourse elicitation tasks (Fergadiotis & Wright, 2011; Stark, 2019). We examined this relationship in our data, collapsed across modalities, prior to running our linear mixed-effects models. Table 11 characterizes the broad effect of prompt category on our measured discourse features.

Table 11. Discourse measures by prompt category

	Expositional (N=92)	Storytelling (N=92)
Word Count		
Mean (SD)	188 (240)	468 (260)
Median [Min, Max]	139 [33.0, 1730]	385 [163, 1710]
Word Length in Number of Letters		
Mean (SD)	3.86 (0.246)	4.03 (0.241)
Median [Min, Max]	3.83 [3.35, 4.60]	4.00 [3.47, 4.61]
Syllable Count		

	Expositional (N=92)	Storytelling (N=92)
Mean (SD)	228 (284)	595 (327)
Median [Min, Max]	171 [44.0, 2050]	499 [213, 2170]
Syllables per Word		
Mean (SD)	1.22 (0.0603)	1.28 (0.0746)
Median [Min, Max]	1.21 [1.11, 1.46]	1.27 [1.17, 1.48]
Word Age of Acquisition		
Mean (SD)	4.40 (0.165)	4.63 (0.112)
Median [Min, Max]	4.39 [4.06, 4.85]	4.62 [4.45, 5.10]
Word Frequency		
Mean (SD)	7420 (978)	6590 (457)
Median [Min, Max]	7440 [4860, 10700]	6540 [5430, 7680]
Word Concreteness		
Mean (SD)	2.84 (0.146)	2.68 (0.0975)
Median [Min, Max]	2.85 [2.52, 3.22]	2.69 [2.47, 2.88]
Phonemes per Word		
Mean (SD)	2.99 (0.162)	3.11 (0.214)
Median [Min, Max]	2.97 [2.72, 3.57]	3.06 [2.73, 3.63]
MTLD		
Mean (SD)	44.2 (13.1)	47.9 (10.6)
Median [Min, Max]	42.6 [20.5, 85.2]	46.4 [29.6, 86.9]
Shannon entropy		
Mean (SD)	3.93 (0.395)	4.51 (0.195)
Median [Min, Max]	3.97 [2.95, 4.97]	4.50 [4.08, 5.03]

Note. Min = Minimum; Max = Maximum; SD = Standard deviation. MTLD = Measure of Textual Lexical Diversity; TTR = Type-Token Ratio.

Lexical diversity (e.g., MTLD) and information density (e.g., Shannon entropy) were decreased in expositional prompts compared to storytelling prompts (Table 10). Compared to the storytelling prompts, the expositional prompts were on average shorter (e.g., decreased word and syllable counts across documents) and contained higher-frequency words.

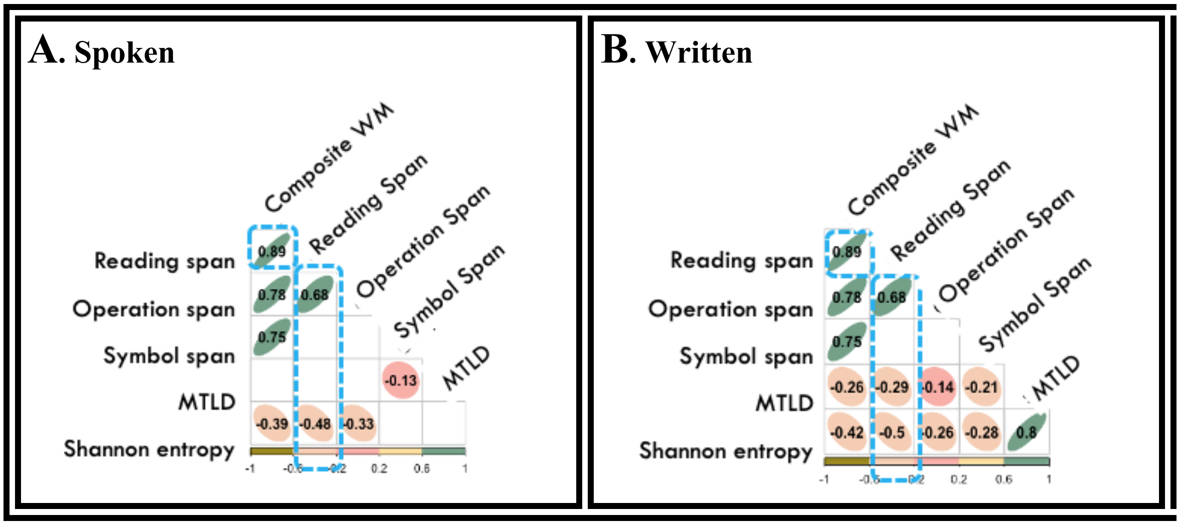
Across both expositional and storytelling prompts, MTLD in written narratives tended to be greater than in spoken narratives ($MTLD_W > MTLD_S$). MTLD was greatest for written storytelling prompts (as expected) and least for spoken expositional narratives (as expected). Prompt-based discrepancies in MTLD were greater in written narratives than in spoken narratives (*mean difference*$_W$ = 5.5; *mean difference*$_S$ = 1.9). Shannon entropy was greater in storytelling vs. expositional prompts, with negligible differences observed by modality within prompt category. Spoken narratives tended to be longer (e.g., Word Count; Syllable Count) than written narratives for both prompt categories. Expositional prompts were shorter than storytelling prompts, regardless of modality. Mean AoA was slightly greater for storytelling vs. expositional prompts and did not vary greatly by modality within prompt category. Mean word frequency was somewhat greater for expositional vs. storytelling narratives, with slightly reduced word frequency observed in written vs. spoken narratives for both expositional and storytelling prompts.

Statistical Outcomes: Linear Mixed-Effects Modeling

In this project, we investigated the relationship between working memory and two primary outcome variables (i.e., information density and lexical diversity), contrasting spoken and written language samples elicited from neurotypical young adults. We predicted that working memory would positively predict lexical diversity (measured via MTLD) and information density (measured via Shannon entropy), with greater effects in spoken vs. written language. As shown in Figure 6, the Reading Complex Span Task was the WM measure most highly correlated with participants' mean MTLD and mean

Shannon entropy at *p>0.05*. Thus, we chose the Reading Complex Span Task to index working memory in our linear mixed-effects models.[4]

Figure 6. Correlation matrix of working memory measures and outcome variables



Note. Only correlations significant at *p> 0.05* are displayed. Shannon.mn = Shannon entropy ; Reading span = Reading Complex Span Task; Operation span = Operation Span Task; Symbol span = Symbol Span Task; MTLD= Measure of Textual Lexical Diversity.

We used a multivariate approach in our respective analyses, predicting lexical diversity (measured via MTLD) and information density (measured via Shannon entropy) using linear mixed-effects models. We included working memory, modality, and working memory*modality as fixed effects, with random effects of prompt category (e.g., expositional vs. storytelling), participant, and prompt. We ran additional linear-mixed effects models to incorporate each of our additional neuropsychological measures as a fixed effect, one at a time; however results of this analysis were not significant (see Appendix B for overview of model outputs). Based on the results of our descriptive analysis (e.g., Table 8), we included prompt (e.g., Broken Window, Cat Rescue,

[4] It is worth mentioning that, among the complex span tasks, the Reading score was also the most highly correlated with the composite WM score. This is perhaps unsurprising when one considers the fact that some degree of verbally-loaded WM processing is required for the participant to hold and manipulate linguistic information online, thereby enabling their successful study participation. There is certainly space for a larger theoretical discussion around the modality (in)dependence of WM; however, this was not the focus of the project at hand and so will not be discussed in detail in this document.

Cinderella, *Snack Attack*) as another random effect in our final models. In the MTLD model, prompt accounted for additional variance otherwise attributed to random effects. Although prompt and other random effects account for vanishingly little variance in the Shannon entropy model, for comparison purposes, we present results from the same set of model inputs for both of our variables of interest in Table 12.

Table 12. Linear mixed-effects model outputs

	Shannon Entropy			**MTLD**		
Predictors	**Estimates**	**CI**	**p**	**Estimates**	**CI**	**p**
(Intercept)	4.31	3.89 – 4.72	<0.001	41.22	37.16 – 45.29	<0.001
Working Memory$_R$	-0.15	-0.26 – -0.04	0.008	-0.44	-4.33 – 3.46	0.826
Modality$_W$	-0.12	-0.18 – -0.06	<0.001	10.39	7.80 – 12.97	<0.001
WM$_R$ × Modality$_W$	0.01	-0.07 – 0.09	0.827	-3.04	-6.28 – 0.20	0.066
Random Effects						
σ^2	0.04			74.65		
τ_{00}	0.03 $_{\text{participant}}$			35.22 $_{\text{participant}}$		
	0.01 $_{\text{prompt}}$			6.36 $_{\text{prompt}}$		
	0.08 $_{\text{prom\_cat}}$			0.35 $_{\text{prom\_cat}}$		
ICC	0.74			0.36		
N	2 $_{\text{prom\_cat}}$			23 $_{\text{participant}}$		
	23 $_{\text{participant}}$			4 $_{\text{prompt}}$		
	4 $_{\text{prompt}}$			2 $_{\text{prom\_cat}}$		
Observations	184			184		
Marginal R^2 / Conditional R^2	0.088 / 0.761			0.193 / 0.483		

Note. WM = working memory; Working Memory$_R$ = Working memory measured by the Reading Complex Span Task; Working Memory$_O$ = Working memory measured by the Operation Complex Span Task; Working Memory$_S$ = Working memory measured by the

Symbol Complex Span Task; prompt = prompt category (e.g., Broken Window, Cat Rescue, *Cinderella*, *Snack Attack*); prom_cat = prompt category (e.g., expositional vs. storytelling).

In the MTLD model, modality_W emerged as the only significant fixed effect (estimate = 10.39, *p<0.001*), with overall model *marginal* R^2 = 19.3%, *conditional* R^2 = 48.3%, *ICC* = 36%. Meanwhile, working memory and modality_W were significant in the Shannon entropy model, with overall model *marginal* R^2 = 8.8%, *conditional* R^2=76.1, *ICC* = 74%. An interaction effect of working memory*modality was not significant in either model, but at $p \cong 0.07$, seems to be approaching significance in the MTLD model.

Canonical Correlation Analysis

Table 13 displays the canonical coefficients yielded for the measures included in the predictor and outcome variable sets by modality, grouped into canonical variates. Canonical variates were significant at $p < 0.05$ using Roy's largest root with F approximation (*df* = 22); F = 4.83 in the spoken model and F=3.55 in the written model. Canonical coefficients represent the relative variance accounted for by each of the included variables and are interpreted similarly to beta weights in regression models or factor loadings in principal component analysis (Iweka & Anthonia, 2018). The spoken model yielded canonical correlation = [0.31, 0.17]. In the written model, canonical correlation = [0.27, 0.05]. It appears that working memory as indexed by the Predictor Set (Table 12) is slightly more correlated with Shannon entropy and MTLD (e.g., the Outcome Set) in spoken (canonical correlate = 31%) than in written language (canonical correlate = 27%). Examination of the second canonical correlation reveals that Shannon entropy and MTLD are dissimilarly predictive of working memory when measured from spoken language (canonical correlation = 17%) vs. written language (canonical correlation = 5%).

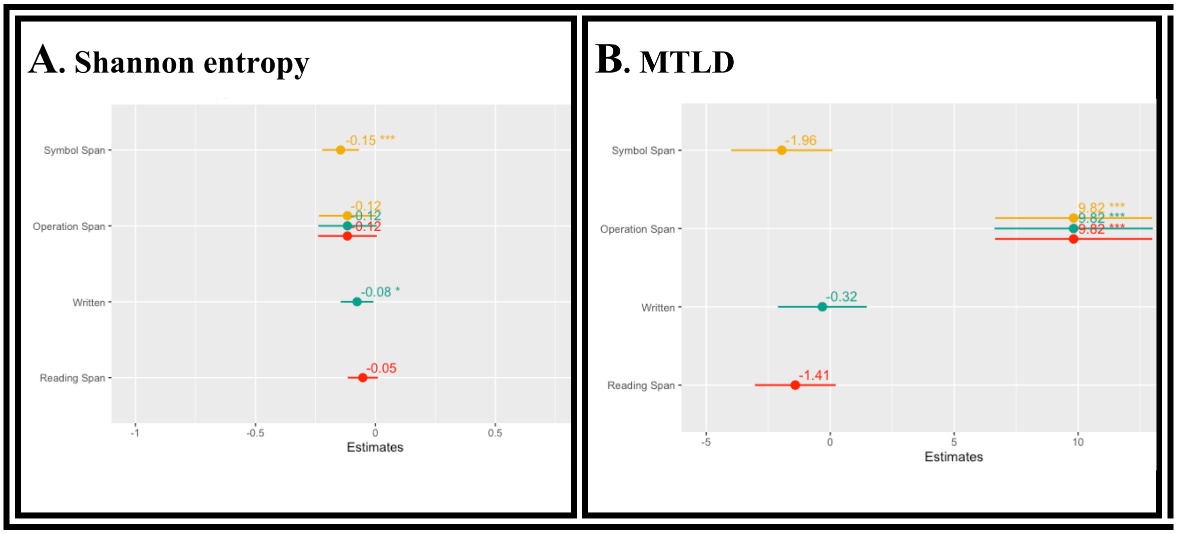
Table 13. Canonical dimensions and associated coefficients from CCA

	Variable Name	Canonical Variate	
		Canonical Variate 1	Canonical Variate 2
Predictor Set	Working Memory$_{R}$		
	Spoken	-1.99	1.24
	Written	-1.47	-0.74
	Working Memory$_{O}$		
	Spoken	-0.38	-1.24
	Written	0.34	-0.20
	Working Memory$_{S}$		
	Spoken	0.37	0.38
	Written	-0.07	1.14
Outcome Set	Shannon entropy		
	Spoken	2.63	-0.24
	Written	1.09	2.53
	MTLD		
	Spoken	-0.07	-0.11
	Written	0.02	-0.10

Note. Working Memory$_{R}$ = Working memory measured by the Reading Complex Span Task; Working Memory$_{O}$ = Working memory measured by the Operation Complex Span Task; Working Memory$_{S}$ = Working memory measured by the Symbol Complex Span Task; MTLD = Measure of Textual Lexical Diversity.

Lastly, we ran simple linear models for each of our individual working memory measures and MTLD and Shannon entropy, respectively. Results are displayed in Figure 7. All beta estimates for predictor variables were negative, and among the three working memory measures, the Reading Complex Span Task was the strongest predictor of both MTLD and Shannon entropy, although this relationship was significant only in the Shannon entropy model. On the whole, beta weights for predictor variables in the Shannon entropy linear models tended to be smaller and more tightly clustered than those observed for MTLD. In the Shannon entropy model, Operation Span was the next strongest predictor (*p > 0.05)*, closely followed by Symbol Span (*p=0.1)*. The opposite results are observed in the MTLD model, where Symbol Span (*p=0.08)* was a stronger predictor than Operation Span. However, none of the working memory measures yielded significant results in predicting MTLD; rather, modality is the only significant variable in this analysis.

Figure 7. Simple linear model outputs



Note. Simple linear models run to predict Shannon entropy (A) and MTLD (B) using each of the three working memory span task scores. MTLD = Measure of Textual Lexical Diversity.

Discussion

In this project, we examined the relationship between working memory, language modality, and discourse features in healthy young adults using a mixture of standard (e.g., Broken Window, Cat Rescue, Cinderella) and novel (e.g., *Snack Attack*) discourse elicitation tasks. We predicted that working memory would positively predict information density and lexical diversity across modalities, with stronger effects in writing than in speaking. Results of linear-mixed effects modeling revealed a negligible role for modality in mediating information density (e.g., Shannon entropy), but strong effects of modality for lexical diversity (e.g., MTLD). To follow, we summarize and interpret the results of our various analyses and consider the theoretical and clinical implications of our findings.

Working Memory and Lexical Diversity

We predicted that working memory (indexed by the Reading Complex Span Task score, or working memory - reading; WM_R) would positively predict lexical diversity (indexed by MTLD) in speaking and in writing; with greater effects observed in the former modality. However, results from a linear mixed-effects model predicting MTLD did not entirely support this prediction (WM_R *model estimate* $= -0.44, p = 0.83$). A similarly negative relationship was observed in correlations of participants mean-aggregated MTLD score and $WM_{R,}$ although results were significant only in written language ($r_{spoken} = -0.05, p = 0.17; r_{written} = -0.29, p < 0.01$). This is of interest because it appears that much of the variability observed in the MTLD model can be attributed to the individual participant ($\tau_{00} = 35.22$), followed by prompt ($\tau_{00} = 6.36$), then prompt category ($\tau_{00} = 0.35$). It appears that individual variability is a stronger predictor of MTLD than contextual factors (e.g., prompt) already known to influence lexical diversity and other discourse measures (Stark, 2019). It is possible that this finding is related to the inherently individualized nature of spoken and written language exposure and acquisition over the course of a human lifespan and manifested among neurotypical individuals in varied patterns of responses to a range of language processing tasks (Hoffman, 2018; Mirman & Graziano, 2012). We observed a strong effect of modality (*model estimate* $= 10.39, p > 0.001$) such that writing tended to elicit a broader range of unique vocabulary words than speaking ($MTLD_W > MTLD_S$). This general finding is consistent with previous reports of relatively greater lexical diversity measured in written vs. spoken language, a phenomenon also observed in descriptive statistics generated from n=93 samples of spoken and n=93 samples of written language in the current study.

Working Memory and Information Density

We indexed information density using Shannon entropy, predicting that with increased WM_R, Shannon entropy would also increase. In linear mixed-effects modeling, WM_R emerged as a significant predictor of Shannon entropy; however, as was the case with MTLD, the observed effect was not in the anticipated direction (*model estimate* = -0.15*, p <0.008*). Similarly negative correlations (significant at *p >0.05*) were observed between participant mean-aggregated Shannon entropy and WM_R in spoken (r^2*= -0.48)* and written (r^2*= -0.5)* language samples. Although modality was a significant predictor (*model estimate*= -0.12, *p < 0.001*), its overall effect on Shannon entropy was relatively small (Shannon entropy_W < Shannon entropy_S). Taken together, these results suggest that Shannon entropy is generally not strongly influenced by language modality. Instead, WM_R and Shannon entropy appear to share a generally similar proportion and degree of variance across language modalities. Thus, we found partial support for our predicted outcome: WM_R predicts Shannon entropy in both spoken and written language samples. However, results were similar across modalities and in the opposite of the anticipated positive direction: decreased working memory appears to predict a slight increase (~2.7%) in Shannon entropy.

Interpretations and Considerations for Future Study

Working Memory Measures and Information Density. In linear mixed-effects modeling, we indexed working memory using the Reading Complex Span Task score, based on correlational analyses among our measures of interest in both spoken and written language. However, there is ongoing debate surrounding the modality specificity of working memory resources (Cowan, 2008; N. Martin et al., 2020; R. C. Martin et al., 2020). We used CCA to examine the overarching effects of working memory on lexical diversity and Shannon entropy, yielding two canonical variates for each modality. We found that the predictor dataset of working memory scores was similarly correlated with the outcome dataset in both spoken and written language (CC_{Spoken} = 31%; $CC_{Written}$= 27%). However, it did not appear that Shannon entropy and MTLD (i.e., the outcome dataset, approximating information density) were equally predictive of the working memory dataset across modalities: CC_{Spoken} = 17%; $CC_{Written}$= 5%. Examining the output of simple linear regression models run for each of our working memory scores, along with the overall pattern of results reported across study analyses, it appears that Shannon entropy is not heavily influenced by language modality. Rather, it seems to be more predictive of general cognitive ability (e.g., WM, processing speed). MTLD, on the other hand varies between spoken and written modalities, as well as by prompt type.

In this project, we indexed working memory using a series of complex span tasks with the goal of accurately characterizing human cognitive processing capabilities in an 'online' assessment. We then assessed the predictive power of working memory on spoken and written discourse measures indexing *information density* (e.g., MTLD, Shannon entropy). Overall, we observed minimal effects of modality on Shannon entropy measured from discourse language samples produced by healthy young adults. Although Shannon entropy is not significantly impacted by language modality; it does appear to index some general cognitive ability. Unlike Shannon entropy, MTLD is linked to salient lexical, phonological, and semantic representations. By nature, lexical and phonological features of language will be impacted by modality. However, Shannon entropy is an index based on simple co-occurrence data. Thus, although MTLD and Shannon entropy are highly correlated and both may be considered indices of *information density*, they accounted for disparate sources of variance in the current analysis. This overarching interpretation warrants further investigation in a future study.

Pragmatics. In designing our experiment, we attempted to control for effects of pragmatic influences on language production processes in order to isolate effects of our predictor variables (e.g., working memory, processing speed, vocabulary knowledge). However, it is likely that we were unable to entirely eliminate pragmatic factors due to: 1) the artificial nature of the experimental task; and 2) persistent effects of pragmatically-induced temporal constraints secondary to generally increased exposure to and use of spoken language in day-to-day contexts. In other words, although the researcher left the room and participants were provided with clear prompt instructions, it is possible that these actions did little to alleviate the pragmatically-induced temporal demands associated with spoken expression. Awareness of pragmatic cues is influenced by theory of mind; that is, one's ability to accurately infer information about the inner state (i.e., knowledge and beliefs) of others (Astington & Jenkins, 1999; Hale & Tager-Flusberg, 2005). We did not measure theory of mind in this cohort of healthy young adults; however, given the limitations described above, it may be informative to include a measure of theory of mind in future studies. Alternatively, examining these relationships in a population with established impairment in theory of mind in a future study could prove informative in refining our understanding of the complex interplay of cognitive systems supporting spoken and written expression.

Conclusion

Considered holistically, the relationship between Shannon entropy and MTLD appears to be complex. They are two highly correlated discourse measures (r_S =.45, p < .0001; r_W = 0.66, p < .001), yet demonstrated differential effects of modality and prompt in the current study (Biber, 2004; Chafe & Tannen, 1987). We attributed the observed differences in these two measures of *information density* to the fact that, while both serve to measure new or unique information, MTLD indexes additional linguistic information (e.g., semantic, lexical). In contrast, Shannon entropy is based on word co-occurrence statistics. Although both measures are indexed along numeric vectors, the representations of words (e.g., ‘word embeddings’) in Shannon entropy are considered *hyperparameters* – that is, they do not correspond to a physically measurable phenomenon. It is possible that this putative additional information contributed to linking MTLD and language modality, whereas the same effects were not observed for Shannon entropy. This interpretation, along with other limitations mentioned previously, warrants exploration in a future study.

CHAPTER 5

GENERAL DISCUSSION

In this project, we sought to characterize the relationship between discourse measures, cognitive function, and effects of language modality (e.g., oral vs. written). We focused our approach on discourse measures linked to working memory (WM) ability (e.g., MTLD and Shannon entropy) via their association with the construct of *idea density* (i.e., *informativity)*. Working within a theoretical framework constructed from convergent, multi-disciplinary evidence (Bryant et al., 2016; Chafe & Tannen, 1987; Kellogg et al., 2013; Nicholas & Brookshire, 1993; Olive & Kellogg, 2002; Shannon, 1950; Shannon & Weaver, 1949; Stark, 2019; Yancheva & Rudzicz, 2016), we proposed that working memory (WM) underpins language processing across modalities, and differential WM demands enacted in speaking vs. writing drive observed differences in discourse measures collected from spoken vs. written data. In Experiment 1, we took a computational approach, using natural language processing methods to explore how language features are differentially distributed in spoken vs. written English, independent of well-known effects of context (e.g., at a baseball game vs. at a wedding) and genre (e.g., expository, storytelling). We predicted that the statistical distribution of language features would significantly vary by language modality. In a large corpus analysis, we found support for our prediction: a supervised machine learning algorithm trained on an 80/20 train/test split achieved >90% accuracy across multiple testing iterations of a binary classification method known as a support vector machine. We thus established initial evidence that some degree of observed variance in discourse measures is likely due to modality alone, in addition to well-known variance attributable to context and genre. In Experiment 2, we expanded on the results of Experiment 1 to propose and test a novel cognitive-processing model of narrative language. We kept a theoretically motivated focus on the role of working memory in receptive vs. expressive processing demands enacted in oral vs. written discourse, predicting that the relationship between *informativity* and objective measures of WM would be attenuated by language modality, with greater effects observed in spoken language. We observed divergent results for our primary measures of *informativity/idea density* (e.g., MTLD and Shannon entropy). MTLD varied by modality and by prompt; however, Shannon entropy was not observed to significantly vary by modality as assessed via linear-mixed effects modeling and canonical correlation analysis. To follow, we discuss the relationship between MTLD and Shannon entropy, two measures of *informativity/idea density* that appear to index differential variance attributable to WM. We contextualize the relationship between Shannon entropy and generative language models, framing our general discussion within the promises and pitfalls associated with the use of such tools, sometimes referred to as "artificial intelligence."

Meaningfulness in Human Language vs. Large Language Models

Statistical learning refers to a process by which repeated exposure to a phenomenon or system builds cumulative knowledge, yielding a stable representation over time (e.g., the phenomenon is 'learned'). Some language scientists argue that statistical learning is a key mechanism driving the acquisition and maintenance of abstract concept knowledge (e.g., truth, joy) stored in semantic memory (Barsalou, 2016; Binder, 2016). Supporting evidence for this claim may be observed in the word frequency effect, referring to the fact that people tend to respond more quickly to a given word if they encounter it more often (Monsell et al., 1989). Similarly, feature-based approaches to the structure and organization of semantic memory suggest that concepts which tend to co-occur tend to be more similar (e.g., are closer together in Euclidean geometric space) than concepts with decreased co-occurrence. Measures of feature similarity in such models may span multiple levels of language processing (e.g., lexical, phrasal). Architects behind generative language models (e.g., OpenAI's *ChatGPT*, Google's *Bard*) attempted to capture word meaning by following this logic, using matrix algebra to estimate the most likely next word in a given linguistic context. In generative language models, these estimates are based on co-occurrence parameters extracted from massive linguistic datasets and therefore, some argue, offer insights into human statistical learning in language acquisition (Contreras Kallens et al., 2023). While the insights drawn from large-scale data analyses are valuable for improving our understanding of the structure and function of language as a whole, it is difficult to extend this application to implementation in a real-world setting (e.g., post-stroke language rehabilitation) given the profound difference in the acquisition and representation of *meaning* in humans versus machines.

Modern computer science was developed in the context of information theory with special attention paid towards the linguistically-derived informativity measure of Shannon entropy. While this approach to computer science was beneficial in that it is scalable, since the philosophy underpinning Shannon entropy was derived from language, it is non-additive. In other words, since language is an emergent, complex system, the sum of the parts (i.e., lexical informativity) does not equal the sum of the whole (i.e., textual informativity). Such non-linear scaling is in fact typical of complex systems, including language (Massip-Bonet et al., 2019).

In computer science, the non-additive nature of informativity indexed across different levels of language processing is putatively solved by using increasingly large language samples to estimate word co-occurrence frequencies. By applying linear algebra, these co-occurrence estimates are transformed into vector representations termed *hyperparameters* (Lantz, 2013). Among the language researchers that consider hyperparameters analogous to feature-based characterizations of semantic memory (i.e., Contreras Kallens et al., 2023), some have attempted to classify the type and degree of semantic relatedness between lexical items (Reilly et al., 2022) by using a mix of hyperparameters and psycholinguistic indices generated from crowd-sourced data (Brysbaert et al., 2014; Kuperman et al., 2012). However, the non-additive nature of *meaning* embedded across various levels of language suggests that it is inappropriate to apply informativity measures (i.e., hyperparameters) derived from aggregated lexical co-occurrence statistics, which in turn are drawn from multiple large samples of lexical data, to examine *meaning* at the level of the individual, be it a word or a human being. Rather, our results suggest that it is more accurate to examine *meaning* (i.e., *informativity/idea density)* not through simple word frequency co-occurrence measures of informativity (i.e., as in Shannon entropy), but rather, to use the seeming 'point of stability' described in the development of MTLD (McCarthy, 2005; McCarthy & Jarvis, 2010) as an example of a latent cue embedded in language. Such a cue, while not detectable to the human eye, is quickly and accurately captured via machine learning methods and based on the results of our experiments, appears informative in refining our understanding of the underlying emergent properties of language and other complex systems (Massip-Bonet et al., 2019).

The ability to quickly and accurately capture discourse measures across various levels of language processing (e.g., lexical vs. textual) using machine learning represents a promising tool for language researchers. In speech-language pathology and other potential areas of clinical application, it is vital that machine learning methods are interpreted and deployed within a theoretically motivated framework. With appropriately meticulous application and interpretation, machine learning methods may prove useful over the long-term in regards to future clinical applications.

Long-Term Potential for Clinical Applications

Critically, machine learning algorithms are capable of detecting even subtle changes in high-dimensional data and have shown promise in accurately predicting disease in a range of populations. The benchmarking language measures generated in this study have the potential to further our understanding of relative impairment in clinical populations (e.g., aphasia, dementia), as well as provide normative data for use in developing cognitive-linguistic screening tools based on discourse measures. Language researchers and clinicians (i.e., speech-language pathologists) have long used discourse analysis to supplement standardized language testing and to evaluate treatment outcomes for PWA (Bryant et al., 2016). In some cases, PWA demonstrate dissociated language impairment in speaking vs. writing (Basso et al., 1978); thus, in a clinical context it is critical to elicit a language sample in both modalities to inform clinical decision making on issues including the severity of language impairment, rehabilitative potential, and approaches to treatment. Such an approach within medical speech-language pathology would align with increasingly popular “precision medicine” approaches to health care. Precision medicine represents a departure from a “one-size-fits-all” approach to treating disease: in precision medicine, health care is flexibly adapted based on an individual’s personal and health histories as well as sociocultural and environmental factors. Health care systems implementing precision medicine frameworks demonstrate improved patient outcomes and reduced operating costs (Alyass et al., 2015; Kasztura et al., 2019). However, future research is needed to fully characterize the potential and pitfalls of using big data to inform individual level decisions surrounding healthcare treatment and rehabilitation. In future research, we plan to refine our characterization of distributional features (i.e., discourse measures) typical of oral and written expression by analyzing language generated from various storytelling prompts widely used in clinical and research-oriented speech-language pathology practices (e.g., the broken window story, the *Cinderella* story). We plan to include clinical populations in future behavioral studies in order to further characterize the relationship between various cognitive systems and discourse measures, including a broader range of lexical diversity indices in our analysis (McCarthy, 2005; McCarthy & Jarvis, 2010). As demonstrated here and in prior research examining the validity of MTLD as a length-invariant index of informativity (i.e., specifically, lexical diversity), different approaches to measuring the same construct can yield heterogenous effects across different levels of analysis. Eventually, the creation of a large database of such freely-available normative language measures has the potential to facilitate speech-language pathologists' screening and evaluation of discourse in both clinical and research practice. By applying machine learning methods to analyze discrepancies typical of discourse features in disordered vs. neurotypical populations, it is possible than additional latent variables embedded in language may be observed, reported, and subsequently used in the early detection of disease. Providing the additional constraints of a theoretically motivated approach may reduce the amount of unaccounted-for variability observed within and across discourse measures indexing various levels of language processing, thereby improving the degree of confidence in a given measure or proposed mechanism of language production and synthesis.

BIBLIOGRAPHY

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods*, *40*(1), 278–289. https://doi.org/10.3758/BRM.40.1.278

Alamargot, D., Plane, S., Lambert, E., & Chesnet, D. (2010). Using eye and pen movements to trace the development of writing expertise: Case studies of a 7th, 9th and 12th grader, graduate student, and professional writer. *Reading and Writing*, *23*(7), 853–888. https://doi.org/10.1007/s11145-009-9191-9

Allen, L., Likens, A. D., & McNamara, D. S. (2019). Writing flexibility in argumentative essays: A multidimensional analysis. *Reading and Writing*, *32*(6), 1607–1634. https://doi.org/10.1007/s11145-018-9921-y

Allen, L., Snow, E. L., & McNamara, D. S. (2016). The narrative waltz: The role of flexibility in writing proficiency. *Journal of Educational Psychology*, *108*(7), 911–924. https://doi.org/10.1037/edu0000109

Altaf, B., Ali, S. S., & Weber, G.-W. (2020). Modeling the relationship between organizational performance and green supply chain practices using canonical correlation analysis. *Wireless Networks*, *26*(8), 5835–5853. https://doi.org/10.1007/s11276-020-02313-3

Alves, R. A., & Limpo, T. (2015). Progress in Written Language Bursts, Pauses, Transcription, and Written Composition Across Schooling. *Scientific Studies of Reading*, *19*(5), 374–391. https://doi.org/10.1080/10888438.2015.1059838

Alves, R. A., Limpo, T., Fidalgo, R., Carvalhais, L., Pereira, L. Á., & Castro, S. L. (2016). The impact of promoting transcription on early text production: Effects on bursts and pauses, levels of written language, and writing performance. *Journal of Educational Psychology*, *108*(5), 665–679. https://doi.org/10.1037/edu0000089

Baddeley, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839. https://doi.org/10.1038/nrn1201

Baddeley, A. D., & Logie, R. H. (1999). Working Memory: The Multiple-Component Model. In A. Miyake & P. Shah (Eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (pp. 28–61). Cambridge University Press; Cambridge Core. https://doi.org/10.1017/CBO9781139174909.005

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., & al, et. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459.

Barsalou, L. W. (2016). On Staying Grounded and Avoiding Quixotic Dead Ends. *Psychonomic Bulletin & Review*, *23*(4), 1122–1142. https://doi.org/10.3758/s13423-016-1028-3

Basso, A., Taborelli, A., & Vignolo, L. A. (1978). Dissociated disorders of speaking and writing in aphasia. *Journal of Neurology, Neurosurgery & Psychiatry*, *41*(6), 556–563. https://doi.org/10/bdqh94

Behrns, I., Wengelin, Å., Broberg, M., & Hartelius, L. (2009). A comparison between written and spoken narratives in aphasia. *Clinical Linguistics & Phonetics*, *23*(7), 507–528. https://doi.org/10.1080/02699200902916129

Bennett, K. P., & Campbell, C. (2000). Support vector machines: Hype or hallelujah? *ACM SIGKDD Explorations Newsletter*, *2*(2), 1–13. https://doi.org/10.1145/380995.380999

Bernton, E., Jacob, P. E., Gerber, M., & Robert, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *81*(2), 235–269. https://doi.org/10.1111/rssb.12312

Berwick, R. (2003). *An Idiot's guide to Support vector machines (SVMs)*. http://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf

Biber, D. (1986). Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings. *Language*, *62*(2), 384. https://doi.org/10.2307/414678

Biber, D. (2004). If you look at ...: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*, *25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Binder, J. R. (2016). In defense of abstract conceptual representations. *Psychonomic Bulletin & Review*, *23*(4), 1096–1108. https://doi.org/10.3758/s13423-015-0909-1

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, *33*(3–4), 130–174. https://doi.org/10.1080/02643294.2016.1147426

Blair, J. R., & Spreen, O. (1989). Predicting premorbid IQ: A revision of the national adult reading test. *Clinical Neuropsychologist*, *3*(2), 129–136. https://doi.org/10/ftsqpv

Blankenship, J. (1974). The influence of mode, sub-mode, and speaker predilection on style. *Speech Monographs*, *41*(2), 85–118. https://doi.org/10.1080/03637757409375826

Brownsett, S. L. E., & Wise, R. J. S. (2010). The Contribution of the Parietal Lobes to Speaking and Writing. *Cerebral Cortex*, *20*(3), 517–523. https://doi.org/10.1093/cercor/bhp120

Brysbaert, M., & Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual Cognition*, *13*(7–8), 992–1011. https://doi.org/10.1080/13506280544000165

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science*, *27*(1), 45–50. https://doi.org/10.1177/0963721417727521

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. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Brysbaert, M., Wijnendaele, I. V., & Deyne, S. D. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, *104*(2), 215–226. https://doi.org/10.1016/S0001-6918(00)00021-4

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, *15*, 17–29.

Chafe, W., & Tannen, D. (1987). The Relation Between Written and Spoken Language. *Ann. Rev. Anthropol.*, *16*, 383–407.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, *51*(4), 661–703. https://doi.org/10.1137/070710111

Cleland, A., & Pickering, M. (2006). Do writing and speaking employ the same syntactic representations? *Journal of Memory and Language*, *54*(2), 185–198. https://doi.org/10.1016/j.jml.2005.10.003

Clifton, A., Reddy, S., Yu, Y., Pappu, A., Rezapour, R., Bonab, H., Eskevich, M., Jones, G., Karlgren, J., Carterette, B., & Jones, R. (2020). *100,000 Podcasts: A Spoken English Document Corpus*. 15.

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large Language Models Demonstrate the Potential of Statistical Learning in Language. *Cognitive Science*, *47*(3), e13256. https://doi.org/10.1111/cogs.13256

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*(2), 163–183. https://doi.org/10.1016/S0160-2896(01)00096-4

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. https://doi.org/10.3758/BF03196772

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, *104*(2), 163–191. https://doi.org/10/dxmphk

Cowan, N. (1998). *Attention and Memory: An Integrated Framework*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195119107.001.0001

Cowan, N. (1999). An Embedded-Processes Model of Working Memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (pp. 62–101). Cambridge University Press; Cambridge Core. https://doi.org/10.1017/CBO9781139174909.006

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research*, *169*, 323–338. https://doi.org/10/fbq7xw

Crossley, S. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, *11*(vol. 11 issue 3), 415–443. https://doi.org/10.17239/jowr-2020.11.03.01

Crossley, S., & Allen, L. K. (2016). *Incorporating learning characteristics into automatic essay scoring models: What individual differences and linguistic features tell us about writing quality*. *8*(2), 19.

Cunningham, A. E., & Stanovich, K. E. (1991). Tracking the unique effects of print exposure in children: Associations with vocabulary, general knowledge, and spelling. *Journal of Educational Psychology*, *83*(2), 264–274. https://doi.org/10.1037/0022-0663.83.2.264

Cunningham, K. T., & Haley, K. L. (2020). Measuring Lexical Diversity for Discourse Analysis in Aphasia: Moving-Average Type–Token Ratio and Word Information Measure. *Journal of Speech, Language, and Hearing Research*, *63*(3), 710–721. https://doi.org/10.1044/2019_JSLHR-19-00226

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Dattalo, P. V. (2014). A Demonstration of Canonical Correlation Analysis with Orthogonal Rotation to Facilitate Interpretation. *Social Work Publications*.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159–190. https://doi.org/10/dg6tmh

Dell, G., & Anderson, N. (2015). *Models of Language Production in Aphasia* (B. MacWhinney & W. OGrady, Eds.; WOS:000684479900026; p. 577). https://doi.org/10.1002/9781118346136

Dunn, D. (2018). *Peabody Picture Vocabulary Test. [Measurement instrument.]* (5th ed.).

Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic Awareness Instruction Helps Children Learn to Read: Evidence From the National Reading Panel's Meta-Analysis. *Reading Research Quarterly*, *36*(3), 250–287. https://doi.org/10/b7fxq3

Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*, *93*(1), 304–316. https://doi.org/10/cjrx9v

Engle, R. W. (2002). Working Memory Capacity as Executive Attention. *Current Directions in Psychological Science*, *11*(1), 19–23. https://doi.org/10/b5qkt3

Epting, L. K., Gallena, E. M., Hicks, S. A., Palmer, E. N., & Weisberg, T. (2013). Read and think before you write: Prewriting time and level of print exposure as factors in writing and revision. *Journal of Writing Research*, *4*(3), 239–259. https://doi.org/10.17239/jowr-2013.04.03.1

Fergadiotis, G., & Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology*, *25*(11), 1414–1430. https://doi.org/10.1080/02687038.2011.603898

Flower, L., & Hayes, J. R. (1981). A Cognitive Process Theory of Writing. *College Composition and Communication*, *32*(4), 365. https://doi.org/10.2307/356600

Fraser, K. C., Lundholm Fors, K., Eckerström, M., Öhman, F., & Kokkinakis, D. (2019). Predicting MCI Status From Multimodal Language Data Using Cascaded Classifiers. *Frontiers in Aging Neuroscience*, *11*, 205. https://doi.org/10.3389/fnagi.2019.00205

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2015). Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's Disease*, *49*(2), 407–422. https://doi.org/10/f72r27

Gilchrist, A. L., Cowan, N., & Naveh-Benjamin, M. (2008). Working memory capacity for spoken sentences decreases with adult ageing: Recall of fewer but not smaller chunks in older adults. *Memory*, *16*(7), 773–787. https://doi.org/10.1080/09658210802261124

Gonzalez, I., Déjean, S., Martin, P., & Baccini, A. (2008). CCA: An R Package to Extend Canonical Correlation Analysis. *Journal of Statistical Software*, *23*(12). https://doi.org/10.18637/jss.v023.i12

Ho, D. Y. F. (1987). Prediction of foreign language skills: A canonical and part canonical correlation study. *Contemporary Educational Psychology*, *12*(2), 119–130. https://doi.org/10.1016/S0361-476X(87)80045-0

Hoffman, P. (2018). An individual differences approach to semantic cognition: Divergent effects of age on representation, retrieval and selection. *Scientific Reports*, *8*(1), 8145. https://doi.org/10.1038/s41598-018-26569-0

Hoffman, P., McClelland, J. L., & Lambon Ralph, M. A. (2018). Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. *Psychological Review*, *125*(3), 293. https://doi.org/10.1037/rev0000094

Iweka, F., & Anthonia, M.-A. (2018). Canonical Correlation Analysis, A Sin Quanon for Multivariant Analysis in Educational Research. *International Journal of Humanities, Social Sciences and Education*, *5*(7). https://doi.org/10.20431/2349-0381.0507013

Jones, D., & Christensen, C. A. (1999). Relationship between automaticity in handwriting and students' ability to generate written text. *Journal of Educational Psychology*, *91*(1), 44–49. https://doi.org/10.1037/0022-0663.91.1.44

Kellogg, R. T. (2007). Are Written and Spoken Recall of Text Equivalent? *The American Journal of Psychology*, *120*(3), 415–428. https://doi.org/10.2307/20445412

Kellogg, R. T., Turner, C. E., Whiteford, A. P., & Mertens, A. (2016). The role of working memory in planning and generating written sentences. *Journal of Writing Research*, *7*(3), 397–416. https://doi.org/10.17239/jowr-2016.07.03.04

Kong, A. P.-H., Whiteside, J., & Bargmann, P. (2016). The Main Concept Analysis: Validation and sensitivity in differentiating discourse produced by unimpaired English speakers from individuals with aphasia and dementia of Alzheimer type. *Logopedics Phoniatrics Vocology*, *41*(3), 129–141. https://doi.org/10/ghc9qd

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences*, *97*(22), 11850–11857. https://doi.org/10/fjng9n

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Lantz, B. (2013). *Machine learning with R: Learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications* (1st ed.). Packt Publ.

Lanzi, A. M., Saylor, A. K., Fromm, D., Liu, H., MacWhinney, B., & Cohen, M. L. (2023). DementiaBank: Theoretical Rationale, Protocol, and Illustrative Analyses. *American Journal of Speech-Language Pathology*, *32*(2), 426–438. https://doi.org/10.1044/2022_AJSLP-22-00281

LeClere, M. J. (2006). Bankruptcy studies and *ad hoc* variable selection: A canonical correlation analysis. *Review of Accounting and Finance*, *5*(4), 410–422. https://doi.org/10.1108/14757700610712462

Lee, J. J., & Chabris, C. F. (2013). General Cognitive Ability and the Psychological Refractory Period: Individual Differences in the Mind's Bottleneck. *Psychological Science*, *24*(7), 1226–1233. https://doi.org/10/f44kj3

Leonard, L. B., Ellis Weismer, S., Miller, C. A., Francis, D. J., Tomblin, J. B., & Kail, R. V. (2007). Speed of Processing, Working Memory, and Language Impairment in Children. *Journal of Speech, Language, and Hearing Research*, *50*(2), 408–428. https://doi.org/10.1044/1092-4388(2007/029)

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2004). Variation in Language and Cohesion across Written and Spoken Registers. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 843–848.

Lustig, C., Hasher, L., & Tonev, S. T. (2006). Distraction as a determinant of processing speed. *Psychonomic Bulletin & Review*, *13*(4), 619–625. https://doi.org/10/bxjtk6

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for Studying Discourse. *Aphasiology*, *25*(11), 1286–1307. https://doi.org/10.1080/02687038.2011.589893

Magezi, D. A. (2015). Linear mixed-effects models for within-participant psychology experiments: An introductory tutorial and free, graphical user interface (LMMgui). *Frontiers in Psychology*, *6*, 2. https://doi.org/10/gjsk65

Mar, R. A., & Rain, M. (2015). Narrative Fiction and Expository Nonfiction Differentially Predict Verbal Ability. *Scientific Studies of Reading*, *19*(6), 419–433. https://doi.org/10.1080/10888438.2015.1069296

Martin, N., Minkina, I., Kohen, F. P., & Kalinyak-Fliszar, M. (2018). Assessment of linguistic and verbal short-term memory components of language abilities in aphasia. *Journal of Neurolinguistics*, *48*, 199–225. https://doi.org/10.1016/j.jneuroling.2018.02.006

Martin, N., Schlesinger, J., Obermeyer, J., Minkina, I., & Rosenberg, S. (2020). Treatment of verbal short-term memory abilities to improve language function in aphasia: A case series treatment study. *Neuropsychological Rehabilitation*, 1–42. https://doi.org/10.1080/09602011.2020.1731554

Martin, R. C., Rapp, B., & Purcell, J. (2020). Domain-Specific Working Memory: Perspectives from Cognitive Neuropsychology. In R. C. Martin, B. Rapp, & J. Purcell, *Working Memory* (pp. 235–281). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0009

Massip-Bonet, A., Bel-Enguix, G., & Bastardas-Boada., Eds. (2019). *Complexity Applications in Language and Communication Sciences*. Springer Nature Switzerland AG.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McNamara, D. S., Crossley, S., & McCarthy, P. M. (2010). Linguistic Features of Writing Quality. *Written Communication*, *27*(1), 57–86. https://doi.org/10.1177/0741088309351547

Meyer, D. (2023). *Support Vector Machines: The interface to libsvm in package e1071.*

Michalke, M. E. (2021). *tm.plugin.koRpus: Full Corpus Support for the "koRpus" Package*.

Mirman, D., & Graziano, K. M. (2012). Individual differences in the strength of taxonomic versus thematic relations. *Journal of Experimental Psychology. General*, *141*(4), 601–609. https://doi.org/10.1037/a0026451

Mitzner, T. L., & Kemper, S. (2003). Oral and Written Language in Late Adulthood: Findings From the Nun Study. *Experimental Aging Research*, *29*(4), 457–474. https://doi.org/10.1080/03610730303698

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, *137*(2), 267–296. https://doi.org/10.1037/a0021890

Monsell, S., Doyle, M. C., & Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, *118*(1), 43–71. https://doi.org/10/cvh489

Mota, N. B., Copelli, M., & Ribeiro, S. (2017). Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *Npj Schizophrenia*, *3*(1), 18. https://doi.org/10/fg4f

Muraki, E. J., & Pexman, P. M. (2021). Simulating semantics: Are individual differences in motor imagery related to sensorimotor effects in language processing? *Journal* *of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10/gnntzw

Nicholas, L. E., & Brookshire, R. H. (1993). A System for Quantifying the Informativeness and Efficiency of the Connected Speech of Adults With Aphasia. *Journal of Speech, Language, and Hearing Research*, *36*(2), 338–350. https://doi.org/10/ggq4bj

Noakes, M. A., Schmitt, A. J., McCallum, E., & Schutte, K. (2019). Speech-to-text assistive technology for the written expression of students with traumatic brain injuries: A single case experimental study. *School Psychology*, *34*(6), 656–664. https://doi.org/10.1037/spq0000316

Olive, T., & Kellogg, R. T. (2002). Concurrent activation of high- and low-level production processes in written composition. *Memory & Cognition*, *30*(4), 594–600. https://doi.org/10.3758/BF03194960

Open American National Corpus. (2015). *American National Corpus Project*. https://anc.org/data/oanc/

Orimaye, S. O., Wong, J. S.-M., Golden, K. J., Wong, C. P., & Soyiri, I. N. (2017). Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics*, *18*(1), 34. https://doi.org/10.1186/s12859-016-1456-0

Panaretos, V. M., & Zemel, Y. (2019). Statistical Aspects of Wasserstein Distances. *Annual Review of Statistics and Its Application*, *6*(1), 405–431. https://doi.org/10/gg649c

Paulsen, J. S., Romero, R., Chan, A., Davis, A. V., Heaton, R. K., & Jeste, D. V. (1996). Impairment of the semantic network in schizophrenia. *Psychiatry Research*, *63*(2), 109–121. https://doi.org/10/b6pg5x

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10/gfshwg

Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2017). The Calgary semantic decision project: Concrete/abstract decision data for 10,000 English words. *Behavior Research Methods*, *49*(2), 407–417. https://doi.org/10/gk6n3c

Pexman, P. M., & Yap, M. J. (2018). Individual differences in semantic processing: Insights from the Calgary semantic decision project. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(7), 1091–1112. https://doi.org/10/gnkq4c

Pilgrim, C., & Hills, T. T. (2021). Bias in Zipf's law estimators. *Scientific Reports*, *11*(1), 17309. https://doi.org/10.1038/s41598-021-96214-w

Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E. and Fosler-Lussier, E. (2007) Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor).

Planton, S., Jucla, M., Roux, F.-E., & Démonet, J.-F. (2013). The “handwriting brain”: A meta-analysis of neuroimaging studies of motor versus orthographic processes. *Cortex*, *49*(10), 2772–2787. https://doi.org/10/f5mnvx

Ravid, D., & Berman, R. A. (2006). Information Density in the Development of Spoken and Written Narratives in English and Hebrew. *Discourse Processes*, *41*(2), 117–149. https://doi.org/10.1207/s15326950dp4102_2

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring Working Memory Capacity With Automated Complex Span Tasks. *European Journal of Psychological Assessment*, *28*(3), 164–171. https://doi.org/10/f33fq9

Richardson, J. D., Dalton, S. G., Greenslade, K. J., Jacks, A., Haley, K. L., & Adams, J. (2021). Main Concept, Sequencing, and Story Grammar Analyses of Cinderella Narratives in a Large Sample of Persons with Aphasia. *Brain Sciences*, *11*(1), 110. https://doi.org/10/gnfrpq

Rinker, T. (2020). *The quantitative discourse analysis ('qdap') R package.*

Rubin, D. L. (1987). Divergence and convergence between oral and written communication. *Topics in Language Disorders*, *7*(4), 1–18.

Salthouse, T. A. (2011). What cognitive abilities are involved in trail-making performance? *Intelligence*, *39*(4), 222–232. https://doi.org/10.1016/j.intell.2011.03.001

Sauerland, M., Krix, A. C., van Kan, N., Glunz, S., & Sak, A. (2014). Speaking is silver, writing is golden? The role of cognitive and social factors in written versus spoken witness accounts. *Memory & Cognition*. https://doi.org/10.3758/s13421-014-0401-6

Schefzik, R., Flesch, J., & Goncalves, A. (2021). Fast identification of differential distributions in single-cell RNA-sequencing data with waddR. *Bioinformatics*, *37*(19), 3204–3211. https://doi.org/10.1093/bioinformatics/btab226

Shannon, C. (1950). *Prediction and Entropy of Printed English*. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x

Shannon, C., & Weaver, W. (1949). *The Mathematical Theory of Communication* (10th ed.). The University of Illinois Press.

Sirts, K., Piguet, O., & Johnson, M. (2017). Idea density for predicting Alzheimer's disease from transcribed speech. *arXiv:1706.04473 [Cs]*. http://arxiv.org/abs/1706.04473

Slobin, D. L. (1997). Mind, code, and text. In *Essays on Language Function and Language Type: Dedicated to T. Givón* (pp. 437–467). John Benjamins Publishing Company; eBook Academic Collection (EBSCOhost). http://libproxy.temple.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=e000xna&AN=429995&site=ehost-live&scope=site

Stanovich, K. E., & West, R. F. (1989). Exposure to Print and Orthographic Processing. *Reading Research Quarterly*, *24*(4), 402–433. https://doi.org/10.2307/747605

Stark, B. C. (2019). A comparison of three discourse elicitation methods in aphasia and age-matched adults: Implications for language assessment and outcome. *American Journal of Speech-Language Pathology*, *28*(3), 1067–1083.

Straka, M., Hajič, J., & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4290–4297. https://www.aclweb.org/anthology/L16-1680

Taylor, J. E., Beith, A., & Sereno, S. C. (2020). *LexOPS: An R package and user interface for the controlled generation of word stimuli*. 11. https://doi.org/10.3758/s13428-020-01389-1

Tombaugh, T. (2004). Trail Making Test A and B: Normative data stratified by age and education. *Archives of Clinical Neuropsychology*, *19*(2), 203–214. https://doi.org/10.1016/S0887-6177(03)00039-8

Troche, J. (2018). Towards a unified model of semantic memory: Validation and theoretical implications of the conceptual feature rating space. *Language, Cognition and Neuroscience*, *33*(6), 698–709. https://doi.org/10.1080/23273798.2017.1408852

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505. https://doi.org/10.3758/BF03192720

Uttl, B. (2002). North American Adult Reading Test: Age Norms, Reliability, and Validity. *Journal of Clinical and Experimental Neuropsychology*, *24*(8), 1123–1137. https://doi.org/10/fdbn9d

Wechsler, D., 1896-1981. (1981). *WAIS-R : Wechsler adult intelligence scale-revised*. https://search.library.wisc.edu/catalog/999605091402121

Wiley, R. W., & Rapp, B. (2019). Statistical analysis in Small-N Designs: Using linear mixed-effects modeling for evaluating intervention effectiveness. *Aphasiology*, *33*(1), 1–30. https://doi.org/10.1080/02687038.2018.1454884

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00433

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *arXiv:1308.5499 [Cs]*. http://arxiv.org/abs/1308.5499

Yancheva, M., & Rudzicz, F. (2016). Vector-space topic models for detecting Alzheimer's disease. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2337–2346. https://doi.org/10.18653/v1/P16-1221

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 53–79. https://doi.org/10/cjn4jn

Zhu, G., Caceres, J.-P., & Salamon, J. (2022). *Filler Word Detection and Classification: A Dataset and Benchmark* (arXiv:2203.15135). arXiv. http://arxiv.org/abs/2203.15135

APPENDIX A

CLEANING FUNCTIONS FOR SPOKEN AND WRITTEN ENGLISH CORPORA

Figure A1: Spotify Podcast Dataset Cleaning Function (Spoken English)

```
## ------- Write Cleaning Function ----
clean_spot <- function(x) {
  x <- tolower(x)
  x <- gsub("\n", " ", x
  x <-  gsub("(\\d)([a-zA-Z]{3,})", "\\1 \\2", x) # Separate any
digit+alpha combination that has 3 or more letters after the number
  x <-  gsub("(\\d)(st|nd|rd|th)", "\\1", x) #omits -rd, -rd from digits
  x <- gsub("[0-9]{1,}", "", x) # remove all digits, regardless of where
they occur
  x <- gsub("(\\<[b-dB-Df-hF-Hj-nJ-Np-tP-Tv-xV-Xz-zZ-Z]{4,}[a-zA-Z]*)",
"", x) # look for a string of non-consonants (a/e/i/o/u/y) of length 4
or greater
     # at the start of a string and replace the entire word (string)
with nothing
  x <- gsub("([a-zA-Z\\d])\\1\\1{1,}", "\\1", x) # remove any
alphanumeric sequence that is three or more of the same consecutive
letter, replace it with one
  x <- gsub("-", " ", x) # replace dash with space
  x <- gsub("`|´", "'", x)  # replaces tick marks with apostrophe for
contractions
  x <-  replace_contraction(x)
  x <- tolower(x) # make replaced contractions lowercase
  x <- gsub("#", " ", x) # remove hash marks
  x <- gsub("°", " ", x) # remove degree sign
  x <- gsub("[[:punct:]]+" , " ", x) # remove punctuation, replace with
space
  x <- gsub("\\b[b-hj-z]\\b{1}", " ", x) # remove alphabetic singletons
except a or i
  x <- x %>% stripWhitespace()
  x <- x %>% stri_remove_empty()
}
```

Figure A2: Corpus of Contemporary American English Cleaning Function (Written English)

```
clean_coca <- function(x) {

  x <- tolower(x)

  x <- gsub("[(].*?[)]", " ", x) # remove annotations

  x <- gsub("[<].*?[>]", " ", x) # remove annotations

  x <- gsub("\n", " ", x)

  x <-  gsub("(\\d)([a-zA-Z]{3,})", "\\1 \\2", x)  # Separate any
digit+alpha combination that has 3 or more letters after the number
```

```
  x <-  gsub("(\\d)(st|nd|rd|th)", "\\1", x)  #omits -rd, -rd from
digits

  x <- gsub("[0-9]{1,}", "", x) # remove all digits, regardless of
where they occur

  x <- gsub("(\\<[b-dB-Df-hF-Hj-nJ-Np-tP-Tv-xV-Xz-zZ-Z]{4,}[a-zA-Z]*)",
"", x) # look for a string of non-consonants (a/e/i/o/u/y) of length 4
or greater

        # at the start of a string and replace the entire word (string)
with nothing

  x <- gsub("([a-zA-Z\\d])\\1\\1{1,}", "\\1", x) # remove any
alphanumeric sequence that is three or more of the same consecutive
letter, replace it with one

  x <- gsub("-", " ", x) # replace dash with space

  x <- gsub("`|´", "'", x)  # replaces tick marks with apostrophe for
contractions

  x <- gsub("n't", " not", x) # CONTRACTIONS1

  x <- gsub("wo ", " will", x) # CONTRACTIONS2

  x <- gsub("'ll", "will", x) # CONTRACTIONS3

  x <- gsub("'ve", "have", x) # CONTRACTIONS4

  x <- gsub("'re", "are", x) # CONTRACTIONS5

  x <- gsub("'m", "am", x) # CONTRACTIONS6

  x <- gsub(" '", "'", x) #remove space from front of floating '

  x <-  replace_contraction(x)

  x <- tolower(x) # make replaced contractions lowercase

  x <- gsub("[[:punct:]]+" , " ", x) # remove punctuation, replace with
space

  x <-  gsub("@", " ", x) #replace @ with space

  x <- gsub("°", " ", x) # remove degree sign

  x <- gsub("\\b[b-hj-z]\\b{1}", " ", x) # remove alphabetic singletons
except a or i, replace with space

  x <- x %>% stripWhitespace()

  x <- x %>% stri_remove_empty()

}
```

APPENDIX B

WASSERSTEIN DISTANCES FOR SPOKEN VS. WRITTEN DISCOURSE MEASURES

The Wasserstein distance is a metric widely used in statistics, computer science, and machine learning to estimate the minimum amount of work it would take to transform one distribution into another (Panaretos & Zemel, 2019; Schefzik et al., 2021). Recent evidence indicates that the Wasserstein distance provides a less-biased maximum likelihood estimate for power-law distributed data (e.g., word frequency) compared to other maximum likelihood estimates (Bernton et al., 2019; Clauset et al., 2009; Panaretos & Zemel, 2019; Pilgrim & Hills, 2021; Schefzik et al., 2021).

An analogy used to illustrate Wasserstein distance (also called the earth mover's distance) is that of people shoveling dirt between two mounds of earth until each mound exactly matches the other. The minimum amount of work expended to transform the dirt piles is indexed by the Wasserstein distance. The equation used to calculate the Wasserstein distance and its decomposition between two distributions F_A and F_B (Panaretos & Zemel, 2019; Schefzik et al., 2021) reads:

$$d := d(F_A, F_B) = \int_0^1 | F_A^{-1}(u) - F_B^{-1}(u)|^2 \mathrm{d}u$$

$$= (\mu_A - \mu_B)^2 + (\sigma_A - \sigma_B)^2 + 2\sigma_A\sigma_B(1 - \rho_{A,B})$$

Here, $\rho_{A,B} \in |0,1|$ measures differences in distribution *shape* (e.g., skewness) using the Pearson correlation coefficient of all points contained in the quantile-quantile plot of F_A and F_B. Differences in *location* are represented by $(\mu_A - \mu_B)^2$, while differences in distribution *size* are measured using $(\sigma_A - \sigma_B)^2$.

Table B1. Wasserstein Distance Decomposition Values for Discourse Measures

	Distance	Location	Size	Shape
MTLD	1770	1680	35.5	55.4
Shannon entropy	0.129 (	0.0950	0.0289	0.00491
Word Age of Acquisition	0.358	0.325	0.0224	0.0108
Word Concreteness	0.0144	0.0112	0.00301	0.000120
Word Frequency	1550000	1540000	460	9390
Number of Letters per Word	0.412	0.393	0.0126	0.00636
Phonemes per Word	0.356	0.337	0.0130	0.00585
Word Count	40100000	2090000	16800000	21300000

Note. MTLD = Measure of Textual Lexical Diversity.

APPENDIX C

ADDITIONAL LINEAR MIXED-EFFECTS MODELS

Table C1. Linear mixed-effects model outputs for additional fixed effects predicting MTLD

	MTLD + Trails			MTLD + NAART			MTLD + ART			MTLD + PPVT		
Predictors	**Estimates**	**CI**	**p**	**Estimates**	**CI**	**p**	**Estimates**	**CI**	**p**	**Estimates**	**CI**	**p**
(Intercept)	41.30	37.25 – 45.35	<0.001	41.27	37.23 – 45.31	<0.001	41.25	37.20 – 45.31	<0.001	41.18	37.11 – 45.25	<0.001
Working Memory$_{R}$	-0.95	-5.12 – 3.22	0.654	-1.24	-5.71 – 3.23	0.586	-0.29	-4.23 – 3.64	0.883	-0.45	-4.35 – 3.44	0.820
Modality$_{W}$	10.39	7.80 – 12.97	<0.001	10.39	7.80 – 12.97	<0.001	10.39	7.80 – 12.97	<0.001	10.39	7.80 – 12.97	<0.001
Working Memory$_{R}$ × Modality$_{W}$	-3.04	-6.28 – 0.20	0.066	-3.04	-6.28 – 0.20	0.066	-3.04	-6.28 – 0.20	0.066	-3.04	-6.28 – 0.20	0.066
Trails	-0.98	-3.96 – 2.00	0.517									
NAART				1.20	-2.17 – 4.57	0.484						
ART							-0.64	-3.52 – 2.24	0.662			
PPVT										0.40	-2.48 – 3.28	0.783
Random Effects												
σ^2	74.65			74.65			74.65			74.65		
τ_{00}	$34.41_{\text{participant}}$ 6.36_{prompt} $0.33_{\text{prom\_cat}}$			$34.28_{\text{participant}}$ 6.36_{prompt} $0.33_{\text{prom\_cat}}$			$34.85_{\text{participant}}$ 6.36_{prompt} $0.34_{\text{prom\_cat}}$			$35.08_{\text{participant}}$ 6.36_{prompt} $0.34_{\text{prom\_cat}}$		
ICC	0.36			0.35			0.36			0.36		
N	$23_{\text{participant}}$ 4_{prompt} $2_{\text{prom\_cat}}$			$23_{\text{participant}}$ 4_{prompt} $2_{\text{prom\_cat}}$			$23_{\text{participant}}$ 4_{prompt} $2_{\text{prom\_cat}}$			$23_{\text{participant}}$ 4_{prompt} $2_{\text{prom\_cat}}$		
Observations	184			184			184			184		

Marginal R^2 / Conditional R^2	0.199 / 0.483	0.200 / 0.483	0.196 / 0.483	0.194 / 0.483

Note. WM = working memory; Working Memory$_R$ = Working memory measured by the Reading Complex Span Task; Working Memory$_O$ = Working memory measured by the Operation Complex Span Task; Working Memory$_S$ = Working memory measured by the Symbol Complex Span Task; prompt = prompt category (e.g., Broken Window, Cat Rescue, *Cinderella*, *Snack Attack*); prom_cat = prompt category (e.g., expositional vs. storytelling).

Table C2. Linear mixed-effects model outputs for additional fixed effects predicting Shannon entropy

	Shannon entropy + Trails			**Shannon entropy + NAART**			**Shannon entropy + ART**			**Shannon entropy + PPVT**		
Predictors	**Estimates**	**CI**	**p**	**Estimates**	**CI**	**p**	**Estimates**	**CI**	**p**	**Estimates**	**CI**	**p**
(Intercept)	4.31	3.89 – 4.72	<0.001	4.31	3.89 – 4.72	<0.001	4.30	3.89 – 4.72	<0.001	4.30	3.88 – 4.71	<0.001
Working Memory$_R$	-0.17	-0.29 – -0.05	0.005	-0.19	-0.31 – -0.06	0.003	-0.15	-0.27 – -0.04	0.007	-0.15	-0.26 – -0.05	0.005
Modality$_W$	-0.12	-0.18 – -0.06	<0.001	-0.12	-0.18 – -0.06	<0.001	-0.12	-0.18 – -0.06	<0.001	-0.12	-0.18 – -0.06	<0.001
Working Memory$_R$ × Modality$_W$	0.01	-0.07 – 0.09	0.827	0.01	-0.07 – 0.09	0.827	0.01	-0.07 – 0.09	0.827	0.01	-0.07 – 0.09	0.827
Trails	-0.04	-0.13 – 0.05	0.360									
NAART				0.06	-0.04 – 0.16	0.224						
ART							0.02	-0.06 – 0.10	0.621			
PPVT										0.07	-0.01	0.086

				– 0.15
Random Effects				
σ^2	0.04	0.04	0.04	0.04
τ_{00}	0.03 $_{participant}$ 0.01 $_{prompt}$ 0.08 $_{prom\_cat}$	0.03 $_{participant}$ 0.01 $_{prompt}$ 0.08 $_{prom\_cat}$	0.03 $_{participant}$ 0.01 $_{prompt}$ 0.08 $_{prom\_cat}$	0.03 $_{participant}$ 0.01 $_{prompt}$ 0.08 $_{prom\_cat}$
ICC	0.74	0.73	0.74	0.73
N	2 $_{prom\_cat}$ 23 $_{participant}$ 4 $_{prompt}$	2 $_{prom\_cat}$ 23 $_{participant}$ 4 $_{prompt}$	2 $_{prom\_cat}$ 23 $_{participant}$ 4 $_{prompt}$	2 $_{prom\_cat}$ 23 $_{participant}$ 4 $_{prompt}$
Observations	184	184	184	184
Marginal R^2 / Conditional R^2	0.096 / 0.761	0.101 / 0.761	0.091 / 0.761	0.112 / 0.761

Note. WM = working memory; Working MemoryR = Working memory measured by the Reading Complex Span Task; Working MemoryO = Working memory measured by the Operation Complex Span Task; Working MemoryS = Working memory measured by the Symbol Complex Span Task; prompt = prompt category (e.g., Broken Window, Cat Rescue, Cinderella, Snack Attack); prom_cat = prompt category (e.g., expositional vs. storytelling).