Author/s:
Alghamdi, Emad Ahmad A

Title:
Automated video difficulty assessment

Date:
2021

Persistent Link:
https://hdl.handle.net/11343/279376

# Automated Video Difficulty Assessment

by

Emad A. Alghamdi

ORCID: 0000-0002-2316-3457

A thesis submitted in total fulfillment for the
degree of Doctor of Philosophy

in the
School of Languages and Linguistics
Faculty of Arts
**THE UNIVERSITY OF MELBOURNE**

July 27, 2021

# *Abstract*

Automated assessment of text difficulty has been recognized as one method for assisting language teachers, textbook publishers, curriculum specialists, test developers, and researchers to make more informed decisions when selecting texts for use in instruction and assessment. While there is a substantial body of work on written and spoken texts, research on videotext difficulty is very scarce. Through a series of studies, the aim of this research program was twofold: to investigate what makes a videotext difficult for language learners and to develop automated measures to help predict difficulty in videotexts. Constructed to be used in this thesis, the Second Language Video Complexity (SLVC) corpus contains 320 academic lectures and 320 government advertisements which were annotated by 322 intermediate language learners.

In Study 1, the relative contribution of verbal complexity to videotext difficulty was examined. The results demonstrated that videotext difficulty was predicted by variation in pitch, lexical frequency and sophistication, and syntactic complexity. Study 2 sought to investigate the impact of visual complexity on learners' perception of videotext difficulty. To this end, innovative computational measures to gauge visual complexity in videotexts were developed and integrated into the Automated Video Analysis (AUVANA) software. The findings of the study suggested that visual complexity contributes to videotext difficulty and their impact is on a par with that of verbal complexity. Moreover, the result of principal component analysis demonstrated that visual complexity is more likely a multifaceted and multidimensional construct, rather than a unitary construct.

While Study 1 and Study 2 looked at verbal and visual complexity independently, Study 3 focused on the integration of multimodal complexity features into ensemble machine learning models. The findings showed that ensemble multimodal models outperformed unimodal models in predicting difficulty in both video genres. Finally, Study 4 sought to develop an unsupervised approach for forecasting video segment difficulty in real-time. Through leveraging more advanced and sophisticated AI algorithms, several neural network models were trained to forecast difficulty in a corpus of 34,363 video segments. Quantitative and qualitative analyses showed that the trained model performed very well in forecasting difficulty in unseen video segments.

Taken together, this thesis makes clear contributions to the investigation of videotext difficulty assessment. In short, the findings of this thesis revealed the usefulness of automated measures for assessing and predicting videotext difficulty. Also, introduced and developed in this thesis new measures of videotext complexity and computational tools for analyzing, computing, and visualizing complexity in videotexts which may help researchers to perform fine-grain analysis of videotext complexity.

# Declaration of Authorship

I declare that this thesis titled, *Automated Video Difficulty Assessment* and the work presented in it are my own. I confirm that:

- The thesis comprises only my original work towards the Doctor of Philosophy except where indicated in the preface;
- due acknowledgement has been made in the text to all other material used; and
- The length of this thesis, exclusive of tables, bibliographies and appendices, is less than 80,000 words.

# Preface

This doctoral thesis contains only original work by the writer, except for the references that have been appropriately acknowledged. Sections of this thesis contain work that has been presented at conferences or appeared in earlier versions in the following publications:

**Alghamdi, E. A.**, Gruba, P., Marsi, A., & Velloso, E. (In Review). The use of lexical complexity for assessing difficulty in instructional videos. *Language Learning and Technology*.

**Alghamdi, E. A.**, Gruba, P., & Velloso, E. (In Review). The relative contribution of verbal complexity to second language video lectures difficulty assessment. *Modern Language Journal*.

**Alghamdi, E. A.**, Velloso, E. & Gruba, P. (2021). AUVANA: An automated video analysis tool for visual complexity. doi: https://doi.org/10.31219/osf .io/kj9hx.

**Alghamdi, E. A.** (2019). Video difficulty assessment. Presented at the Human-Computer Interaction Seminars. *School of Computing and Information Systems*, The University of Melbourne, Melbourne, Australia.

**Alghamdi, E. A.** (2018). Video complexity and language learners: Making the match. *The Third Saudi Scientific Symbolism*, University of Sydney, Sydney, Australia.

**Alghamdi, E. A.**, Otaif, F., & Gruba, P. a. (2018). Designing a video playing interface for second language learners. In Ascilite (pp. 298–302). doi: 10.1111/j.1365-2729.2011.00476.x

# *Acknowledgements*

I am greatly indebted to numerous individuals. First and foremost, I am sincerely and heartily grateful to my supervisors Dr. Paul Gruba and Dr. Eduardo Velloso for their unwavering support and encouragement, and most importantly for giving me the freedom to explore and experiment with new ideas. My appreciation also goes to my former committee Chair Prof Tim McNamara whose comments on the initial writing of the thesis was so invaluable. My deepest gratitude also goes to my current committee Chair Dr. Carsten Rover for his understanding and encouragement.

During my study, I have met many great people. I wish to express my appreciation to Prof Rachel Fensham for supporting me during my almost two-year residency in the Digital Studio. I would like also to thank my mentors, Dr. Daniel Russo-Batterham and Kim Doyle, for sharing with me some great data analytic tips and for introducing me to the field of computational social sciences during my internship at the Melbourne Data Analytics Platform (MDAP). I am so grateful to Prof. Frank Vetere, the director of the Microsoft Research Centre for Social Natural User Interfaces, for discussing with me some other possible ways for doing my research and for introducing me to Dr. Eduardo.

My Ph.D. experience would not be enjoyable without friends with whom I shared moments of excitement and frustration. Form the School of Linguistics and Applied Linguistics, I would like to thank Abdelaziz Alsharani, Hanan Almalki, and Xiaofang Yao. I also like to thank my colleagues in the Human Interaction Group at the School of Computer Science, especially Dr. Namrata Srivastava, Henrietta Lyons, Ebrahim Babaei, Brandon Syiem, Qiushi Zhou, Anam Khan, Gabriele Marini, and Dr. Joshua Newn.

I am also indebted to my wife Maha and my two kids, Lara and Omar, for their forbearance, especially during the long months of the COVID-19 pandemic and the harsh lockdown in Melbourne. I am thankful for the sacrifices that you made to see this thesis to completion.

Virtually, I am so thankful for many great YouTubers, e.g., 3Blue1Brown, StatQuest, Yannic Kilcher, and Two Minute Papers, to name a few, whose video tutorials taught me a new concept or showed me how to go about solving a bug in my code. What a great time to do a Ph.D.!

Finally, I would also like to thank my sponsor, King Abdulaziz University, and all the students who participated in this investigation.

# *Dedication*

*To my parents*

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| **AI** | **A**rtificial **I**ntelligence |
| **AL** | **A**cademic **L**ecture |
| **EFL** | **E**nglish as a **F**oreign **L**anguage |
| **ESL** | **E**nglish as a **S**econd **L**anguage |
| **GA** | **G**overnment **A**dvertisement |
| **RNNs** | **R**ecurrent **N**euron **N**etworks |
| **SVM** | **S**upport **V**ector **M**achines |
| **SLVC** | **S**econd **L**anguage **V**ideo **C**omplexity corpus |
| **LSTM** | **L**ong **S**hort **T**erm **M**emeory |
| **NLP** | **N**atural **L**anguage **P**rocessing |
| **ARA** | **A**utomated **R**eadability **A**ssessemnt |
| **MML** | **M**ultimodal **M**achine **L**earning |
| **CEFR** | **C**ommon **E**uropean **F**ramework of **R**eference for **L**anguages |
| **SBD** | **S**hot **B**oundary **D**etection |

# Chapter 1

# Introduction

What makes a video difficult for language learners is an enigma. One may contend that since language learners are in the process of learning a new language, many of the difficulties they encounter while watching digital videos ought to be language-related. Although there is undoubtedly some truth to this, video is fundamentally a dynamic, multimodally rich type of text. In ways that are far from being straightforward, language blends in and interweaves with other expressive modalities, inter alia, music, sounds, gestures, facial expressions, scientific symbols, and visual imagery to enhance, extend or contradict meanings; while also possible these modalities convey meanings by themselves, with no help of language (Kress, 2010; Van Leeuwen, 2004). When watching a video, language learners not only need to cope with difficulties in language but also process and keep track of other modalities, establish references to and associations between them, and finally integrate their meanings onto a single coherent mental representation (Paivio, 2007). An intriguing question then arises; how important are these different modalities for language learners' understanding of videos and what modalities or an ensemble of modalities boost or hinder their understanding.

Within the conceptualization of language learning as a semiotic learning process and for which second language (L2) learners need to draw upon a wide range of semiotic resources for producing and comprehending texts (Douglas Fir Group, 2016), watching videos is conceivably a great exercise for language learners to be exposed to socially and multimodally rich semiotic resources. Learning language through watching videos, as opposed to listening to audio texts, resembles everyday listening contexts (Rubin, 1995) and thus is more ecologically valid (Guichon & McLornan, 2008). However, for multimodal language learning to take place, video difficulty needs to be at the right level; not too

challenging to halt language learning or too easy to disengage and demotivate learners (R. Bjork, 1994; Vygotsky, 1978). Taking this alongside the fact that videos have become commonplace in language education (Cross, 2018) and yet there exists no scientific and reliable measures for assessing their difficulty, the primary goal of this thesis is to first investigate what makes a video difficult for language learners, and secondly to develop automated computational models and tools for video difficulty assessment.

Before delving into further discussion of video difficulty, it is helpful to describe what a video is. If we begin to think about how we might want to describe and study videos, it becomes apparent that we can only say limited things about them as media in and of themselves. A dictionary definition of video describes it as "a digital recording of an image or set of images (such as a movie or animation)" (Merriam-Webster, 2021a). A more helpful view of video conceives it as a text—or a videotext (Joiner, 1990). Much like written and spoken texts, videotexts are designed to fulfill a communicative function. A professionally crafted videotext, for example, is meticulously planned, shotted, and edited. Gruba (2004) remarked that by specifically focusing on textual features or elements of videotexts, applied linguists can examine how a videotext generates meaning, engages or disengages a language learner.

In the rest of this introductory chapter, I discuss text difficulty assessment. Then, I motivate the need for investigating videotext difficulty and complexity for use in contemporary listening assessment, teaching, and research. Following that, I introduce the analytical approach I have chosen for investigating and predicting videotext difficulty. Finally, the chapter concludes with an outline of the remaining chapters.

## 1.1 Text Difficulty Assessment

Text difficulty assessment concerns how to accurately assess and predict difficulty in texts as perceived by an individual or a group of individuals. While documented concerns about comprehensibility can be traced back to the classical rhetoric of Plato and Aristotle (Lorge, 1944), a more systematic, scientific approach to the study of text readability only started in the early 1920s. Beginning in earnest with the work of Lively and Pressey (1923), the impetus behind the first readability formula was genuinely practical: how to select appropriate texts for readers with different reading abilities. Over the years, readability researchers have devised numerous metrics or formulas for estimating reading difficulty. While traditional readability formulas generally rely on superficial features of

texts, such as sentence length and word count, contemporary approaches to text difficulty assessment utilize more sophisticated indices to tap into multiple dimensions of text.

In this thesis, I investigate complexity in videotexts—an understudied type of texts—and develop computational models and tools for predicting videotext difficulty. In the following section, I highlight why we need to understand videotext difficulty and how measures of videotext difficulty may contribute to the development of L2 teaching, theory, and research.

## 1.2 Motivations for this Thesis

A key motivation for developing automated videotext difficulty measures is their use in selecting videotexts for language learning, assessment, and research, akin to those already developed for written and spoken texts. While the use of videotexts, e.g., DVDs, podcasts, films, and documentaries, has an extensive history in language teaching and learning (see, e.g., Rubin, 1995; Vanderplank, 2010), it is the recent advancement of video technologies that makes watching videos a principal means for language learning in and out of the classroom (Cross, 2018). Broth, Laurier, and Mondada (2014) noted that "video has come to a point of ubiquity and commonplaceness in the more affluent regions of the world, a reminiscent of the spread of paper and pens". Indeed, on YouTube alone, over one billion hours of materials are watched per day across more than 100 countries in 80 different languages (*YouTube in numbers*, 2021).

Early proponents of the use of videotexts in language teaching saw an opportunity to expose learners to an authentic language situation, one that resembles real-life (Guichon & McLornan, 2008). They asserted that learners not only learn grammatical structures and new vocabulary, but also some sociocultural norms and values of the target language. However, for optimal videotext learning, a healthy balance between videotext difficulty and the learner's ability should be maintained. When a learner is presented with a videotext that exceeds his or her own current ability, learning is more likely to halt (e.g., Vygotsky's (1978) learning theory). An automated system of videotext difficulty can help in selecting videotext for use in language teaching and learning. Other potential applications for automated videotext complexity analysis are in computer-assisted language applications, intelligent tutoring systems, search engines and web applications. Automated videotext complexity analysis may also have potential in modifying original videotext to

make it more accessible for a specific target audience, e.g., language learners and people with disability.

Of equal importance is the use of automated videotext analysis for second language video research. In the last two decades, videotext has received growing attention, as attested in several Special Issues (e.g., Perez & Rodgers, 2019; Peters & Muñoz, 2020), and numerous research studies (see Chapter 3 for a scoping review). Extant research on L2 videotext has mainly concentrated on the use of videotext, but little research has been done on videotext itself. I found this slightly odd because videotext difficulty impacts learners' understanding of videotext and interferes with whatever learners do with it. Accordingly, I believe that not controlling for videotext difficulty in L2 research, regardless of the objective of the study and language proficiency of the participants, may lead to conflicting results, or even worse, false conclusions. A computational measure of videotext difficulty can therefore assist researchers in choosing the most appropriate video materials for their study and make results across other different studies comparable.

Lastly, and perhaps most importantly, investigating what makes videotext difficult for language learners might lead to new insights and discoveries that can help L2 researchers to discover how language learners process, understand, and learn from dynamic audiovisual texts. This eventually puts us on the path of developing a well grounded conceptual understanding of videotext difficulty for L2 learning, which can then be used to inform pedagogical interventions.

## 1.3 Research Approach and Scope

The value of understanding videotext difficulty is hardly controversial, although just how videotext difficulty is to be theoretically conceptualized and empirically investigated is less clear. While comprehension research has primarily focused on printed texts (Mcnamara & Magliano, 2009), theoretical work on sequential visual narrative (e.g., films and comics) is still in its nascent stages (e.g., Aryadoust, 2019; Cohn, 2013, 2016b, 2020; Kendeou et al., 2019; Loschky, Hutson, Smith, Smith, & Magliano, 2018; Loschky, Larson, Smith, & Magliano, 2020; Schnotz, 2013). Therefore, in this thesis, I have taken an evidence-based approach for modeling videotext difficulty. In order to say a videotext feature contributes to videotext difficulty, there must be some empirical evidence supporting that claim. Videotext has been the subject of research across various disciplines, and many claims have been made. In examining empirical evidence, the origin of the

research discipline and the theory supported bear less importance. What matters is the variability of evidence supporting a particular claim. However, a claim that is not based on any theoretical ground is merely idle speculation. Therefore, resorting to and integrating views from well-established theories in areas such as listening, discourse comprehension, visual perception, and attention is necessary to help see the bigger picture.

Building upon theoretically driven complexity features and through leveraging recent advances in the field of machine learning and Artificial Intelligence (AI), this thesis seeks to develop computational models of videotext difficulty assessment in L2. Particularly, videotext complexity assessment is perceived as a supervised machine learning task. To this end, the Second Language Videotext Complexity (SLVC) corpus was constructed. The SLVC is the first annotated videotext corpus for use in videotext complexity and difficulty research. The corpus contains 320 academic lectures (AL) and 320 government advertisements (GA), which were all rated for their difficulty by 322 intermediate English as a Foreign Language (EFL) language learners in a Saudi university.

Necessarily, the scope of this project must be constrained in several ways. The first constraint is in the selection of videotext genres. Two videotext genres, academic lectures and government advertisements, were the only genres investigated in this thesis. The second constraint was in the selection of participants for the study. Only male Saudi students were recruited. The choice of this sample was first due to convenience, and second for ecological reasons. I am more familiar with language learning and teaching in the Saudi higher education system than any other context. Admittedly, the sample would be more representative if female learners were included, but due to the policy of gender segregation in the Saudi education system, it was difficult to recruit female learners. The second reason for limiting the sample to only Saudi college students is that measures of text difficulty should be developed for and validated by a specific target of learners, so that the influence of factors such as sociocultural differences are eliminated. And once developed, text difficulty measures should be used for the same population. As such, the outcomes of this project may not generalize well to other contexts and populations of language learners. The third constraint of the thesis is the mere focus on videotext complexity features that can be automatically extracted and computed, and less attention was paid to high-level qualitative aspects of videotexts such as levels of meaning, structure, and cohesion.

These limitations notwithstanding, this thesis make clear contributions to the fields of automated text difficulty assessment and applied linguistics. This research program was

the first investigation into the impact of videotext complexity on the perception of videotext difficulty. The findings of this thesis provide invaluable insights on how different dimensions of videotext complexity contribute to videotext difficulty and in turn to the degradation of L2 videotext comprehension. In particular, the results demonstrated that both verbal and visual complexities contribute to L2 learners' perception of videotext difficulty. This finding highlights the important role of visual complexity in L2 videotext understanding. Also introduced in this thesis a large annotated corpus of videotexts, new measures and automated software for assessing difficulty in videotexts.

As an applied linguist, I envision the transformation of L2 video research, if not revolutionized, through the development of intelligent tools to complement video research and instruction with computerized affordances that are otherwise unavailable or extremely time and labor-intensive. This thesis illustrates my vision.

## 1.4 Outline of the Thesis

Following this introductory chapter, Chapter 2 reviews previous work on the computational assessment of text difficulty. The chapter is composed of two major sections. The first section provides an overview of text readability research and provides a taxonomy of computational measures of text readability. It also reviews the relatively small, yet growing body of research on written and spoken text difficulty assessment in second language. The second section discusses key shortcomings of automated tools for text difficulty assessment and provides six principles for improving and generating a more reliable assessment of text difficulty.

In Chapter 3, I turn to videotext difficulty. In the first part of the chapter, I review prominent as well as recent theoretical frameworks of visual narrative understanding, and then introduce a preliminary conceptual framework for understanding videotext difficulty within the context of independent video-based language learning. Following that, I venture into collecting and synthesizing empirical evidence of verbal and visual complexity from diverse fields of research.

Chapter 4 describes and discusses key concepts in machine learning that are relevant for developing automated videotext difficulty models. The chapter also outlines the machine learning pipeline implemented in this thesis to develop computational models of videotext difficulty.

In Chapter 5, I describe the process of constructing and annotating the Second Language Videotext Complexity (SLVC) corpus. To the best of my knowledge, the SLVC corpus is the first specialized video corpus for investigating and modeling videotext complexity and difficulty in a second language. The corpus is composed of 320 academic lectures and 320 government advertisement clips. The videotexts were rated for their difficulty by 322 EFL learners, who were all at the B1-level as determined by the Cambridge English proficiency placement test. This chapter also reports the results of validity and reliability analyses on the participants' ratings, followed by a preliminary analysis of the participants' judgement of videotext difficulty.

Chapter 6 reports the first study of this thesis, which investigates the relationship between a wide range of verbal complexity features and videotext difficulty. The chapter first describes how the verbal complexity features were extracted and computed. Next, it reports the findings of regression models, aimed at explaining and predicting videotext difficulty using verbal complexity features. Study 1 also investigated the impact of video genre on the predictive accuracy of videotext difficulty assessment models.

In Study 2 reported in Chapter 7, I investigate the contribution of visual complexity to videotext difficulty. Unlike verbal complexity, little work has investigated the role of visual complexity on language learners' perception of videotext difficulty. To fill this gap, I propose new measures of visual complexity. The proposed measures were subject to factor analysis (using Principle Component Analysis) to examine if the features characterize some latent structures. The chapter also introduces the Automated Video Complexity Analysis (AUVANA), an open-sourced software I have developed to extract, compute, and visualize visual complexity.

Chapter 8 reports on Study 3, which investigates whether combining verbal and visual complexity features improves the prediction of videotext difficulty. Using different configurations of machine learning algorithms and multimodal complexity features, I developed several ensemble learning models that utilize techniques such as averaging, stacking, boosting, and bagging.

Motivated by the prospect of developing intelligent computer assisted language learning systems, I propose an approach that leverages deep learning for developing a real-time video difficulty forecasting system in Chapter 9. The approach allows for forecasting difficulty of short video segments (5 seconds) and hence offers new possibilities in developing adaptive language learning and assessment systems.

Finally, in Chapter 10 I summarize the key findings of the four studies included in this thesis. I also outline the overall contributions and limitations of the thesis. Finally, I end this thesis with an agenda for future research in automated videotext difficulty assessment.

# Chapter 2

# Computational Assessment of Text Difficulty

In Chapter 1, I contextualized the research and outlined its aim and scope. In this chapter, I review traditional and contemporary computational approaches for text readability assessment. I do so because the field of readability has a rich and extensive history and methods of readability assessment can principally be adapted to the analysis of any type of text, including videotexts. Then, I review recent research on the use of text complexity analysis for gauging difficulty in L2 written and spoken texts. In the second half of the chapter, I discuss key major issues and concerns in text difficulty assessment research, and I propose six principles for a more accurate and reliable assessment of text difficulty.

## 2.1 Text Readability: An Overview

In essence, text readability refers to the ease with which a text can be read and understood (Dale & Chall, 1949). With a primary focus on text complexity, readability researchers have come to discover, through observation or experimentation, some characteristics of text make reading more difficult. These elements are manifold and may include, for example, semantic difficulty, complex syntactic structures, but also the lack of supporting graphics and illustration, and myriad issues concerning text legibility and typography, e.g., font type, line height, color, character spacing, and text formatting. A broader and more inclusive conceptualization of readability may include the role of readers and

activity—that is, besides text complexity, readability is also a function of the reader's resources and the type of task they are performing with the text (NGACB & CCSSO, 2010). In this triadic interaction, readers indubitably play a vital role and their comprehension skills, motivation, interests, and educational and social backgrounds, do all contribute to the ease or difficulty of texts. Moreover, what type of activity and how cognitively demanding it is for a reader would also weigh on the perception of text difficulty. That being said, however, automated text readability or difficulty assessment tools focus more on textual factors and less on reader and task factors.

With the advent of increasingly more sophisticated NLP methods along with advances in our understanding of the various cognitive processes underpinning text comprehension, the field of automated text difficulty assessment has made remarkable progress in the last two decades. Contemporary approaches to readability assessment typically involve hundreds of complexity features that tap into different dimensions of text comprehension (Collins-Thompson, 2014). The field has also benefited from the contributions of researchers working in diverse fields, including education, linguistics, cognitive science, discourse processing, computer science, and psychology, who all made strides in devising new methods for estimating and predicting the difficulty of texts for use in different settings (Benjamin, 2012). Reviews of early traditional readability formulas are to be found in Chall (1958) and Zakaluk and Samuels (1988), whereas more recent reviews are to be found in Mesmer (2008), Benjamin (2012), and Collins-Thompson (2014).

Taking a computational linguistics perspective, Collins-Thompson (2014) surveyed computational text readability assessment, focusing largely on algorithms used in the development of computational tools. The survey highlighted the transition from general-purpose readability formulas developed on a small corpus of texts to machine learning-based systems trained on large corpora and rich text features that capture multiple dimensions of text complexity. Benjamin (2012) classified methods of text difficulty assessment into three types: traditional methods (or readability formulas), cognitively inspired measures, and methods based on the use of statistical language modeling. However, since the publication of the above surveys, the field has evolved. In the last two years, a new approach based on deep learning seems to be gaining currency (e.g., Deutsch, Jasbi, & Shieber, 2020; Filighera, Steuer, & Rensing, 2019; Martinc, Pollak, & Robnik-Šikonja, 2019; Nadeem & Ostendorf, 2018).

In the following section, I provide a comprehensive, yet concise, overview of classical and contemporary approaches to text difficulty assessment and outline their key capabilities

and shortcomings. These approaches are categorized into four groups: readability formulas, cognitively inspired approaches, machine learning approaches, and deep learning approaches. There are some similarities as well as differences among these approaches concerning two key factors: types of complexity features (hand-crafted features or learned representations) and type of algorithms (simple, advanced, or neural networks). This relationship is depicted in Figure 2.1

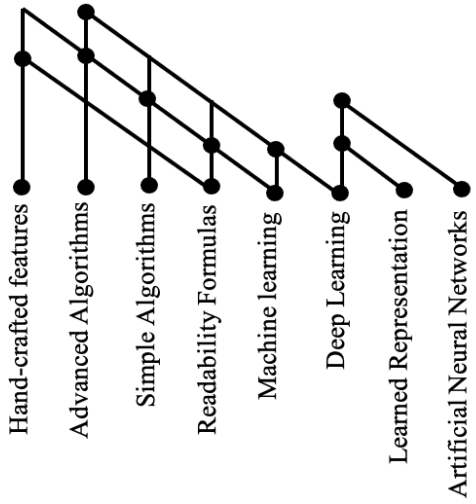



FIGURE 2.1: The Relationship between Complexity Features, Algorithms and Difficulty Assessment Approaches

## 2.1.1 Text Difficulty Assessment Approaches

### 2.1.1.1 Traditional Readability Formulas

Very simply, readability formulas are regression equations in which the dependent variable (what we know and want to predict) is some kind of score reflecting text difficulty for a specific population of readers and the independent or predictor variables are two or more measurable aspects of texts, such as the number of syllables per word and the number of words per sentence (Benjamin, 2012). Once a readability formula is developed, it can be used to predict the reading difficulty of other texts. Though there were some earlier attempts to statistically analyze literature texts (Sherman, 1893), the development of the first readability formula is often credited to Lively and Pressey (Lively & Pressey, 1923). Using Thorndike's vocabulary list of the most common 10,000 words (Thorndike, 1921), Lively and Pressey developed a readability formula to estimate the 'vocabulary burden' in technical textbooks. At that time, vocabulary knowledge was closely associated with the ability to read complex texts and believed to reflect one's verbal intelligence.

Later in 1928, Vogel and Washburne published a landmark work in readability study which had a significant impact on the later development of readability formulas. Unlike Thorndike (1921) and Lively and Pressey (1923), in developing their formula Vogel and Washburne (1928) did not only consider vocabulary difficulty but also various other factors related to sentence structure, parts of speech, paragraph construction, general structure, and physical makeup of the book. They also were the first to use a criterion based on an independent evaluation of text difficulty, children reading scores. Their work encouraged other readability researchers to go beyond vocabulary difficulty as a sole predictor of text readability to include various aspects of text variability.

Beginning in the 1930s, some readability scholars began to focus on adult readers with limited education (e.g., Dale & Tyler, 1934; Ojemann, 1933; Waples & Tyler, 1931). This shift towards adult readers, who were then considered either literate or illiterate, was partially stimulated by the USA government policy to enormously support adult education during the Great Depression (DuBay, 2004). Literacy scholars were interested in understanding why adults readers, especially those with limited schooling, were unable to find materials that meet their interest and reading ability (Dale & Tyler, 1934; Waples & Tyler, 1931). Early surveys of adult reading literacy suggested that the average adult reader in the United States had a limited reading ability. In a comprehensive two-year study, Waples and Tyler (1931) interviewed a group of adult readers and asked them about their interest in books, what they read or do not read, and why. Waples and Tyler (1931) found out that although many adult readers had a keen interest in reading and expanding their knowledge, the books they are interested in were beyond their reading ability. Bailin and Grafstein (2016) noted that Waples and Tyler's study was important because it emphasized finding appropriate reading materials for adults with varying levels of reading ability, akin to what readability formulas had done for children.

The 1940s witnessed the publication of more readability formulas tailored for adult readers; most notable works were those by Flesch (1943, 1948, 1949) and Dale and Chall (1948, 1949). Rudolf Flesch developed his first readability formula in his 1943 Ph.D. dissertation which was a response to what Flesch believed key limitations in earlier readability formulas for adult readers. Following his Ph.D. work, Flesch led a successful career in readability consultation and authored several readability guidelines including his 1946 book, *The Art of Plain Talk*, and his 1949 book, *The Art of Readable Writing*. Flesch was keenly interested in making his formulas accessible to a wide audience. To simplify his formulas, Flesch did not use a reference vocabulary list but rely instead on affixes count as a proxy of word difficulty. The simplicity of his formula made readability formulas

very popular among many textbook publishers and news agencies who were interested in increasing their readership (DuBay, 2004).

Dale and Chall (1948) believed vocabulary difficulty is an important predictor of text readability and developed a formula that puts an emphasis again on vocabulary knowledge. Unsatisfied with previous vocabulary lists, Dale and Chall (1948) created their own list which consisted of 3000 easy words. The formula incorporates vocabulary difficulty and sentence length as only two indices of reading difficultly. Several later studies provided empirical support for the formula and since then the two factors—word difficulty and sentence complexity—were central components in readability formulas developed in the next three decades (DuBay, 2004). With reading research supporting their use in selecting reading materials, the development of readability formulas gained momentum and by the 1980s there were more than 200 readability formulas and thousands of research around their validity and applications (DuBay, 2004).

Traditional readability formulas as well as their modern successors are similar in their approach to computing text readability. To estimate text readability, traditional formulas rely on calculating units, such as the number of syllables per word for a word's semantic difficulty and the length of sentence for syntactic difficulty (Collins-Thompson, 2014). Texts that contain longer sentences and unfamiliar vocabulary would be considered more difficult or less readable than texts that contain shorter sentences and familiar words (Benjamin, 2012). For instance, the Gunning Fog index (GFI: Gunning, 1952) estimates the years of formal education a person needs to understand the text on the first reading. It is calculated with the following equation:

$$GFI = 0.4(\frac{totalWords}{totalSentences} + 100(\frac{longWords}{totalSentences}), \tag{2.1}$$

where *longWords* are words longer than 7 characters. Higher values of the index indicate lower readability.

The Flesch-Kincaid Grade Level (FKGL: Kincaid, Fishburne Jr, Rogers, & Chissom, 1975) computes readability using the following regression equation:

$$FKGL = 0.39(\frac{totalWords}{totalSentences}) + 11.8(\frac{totalSyllables}{totalWords}) - 15.59 \tag{2.2}$$

Flesch Reading Ease (FRE: Kincaid et al., 1975) assigns higher values to more readable texts and lower values to more difficult texts. It is calculated in the following way:

$$FRE = 206.835 - 1.015(\frac{totalWords}{totalSentences}) - 84.6(\frac{totalSyllables}{totalWords}) \tag{2.3}$$

More modern readability formulas also rely on some indices of semantic difficulty and syntactic complexity to estimate readability in texts (e.g., Chall & Dale, 1995; J. A. Stenner, 1996). For example, the Revised Dale-Chall formula was reformed to include Dale's list of 3000 unfamiliar words to 80% of American fourth-graders (Chall & Dale, 1995).

Since their early days, readability formulas have been under attack but it was in the 1980s the level of criticism reached its climax. In conjunction with the cognitive turn in psychology in the 1980s (Gardner, 1987), researchers shifted their attention away from predicting text difficulty to focus on text readers themselves; how readers of different cognitive abilities can read texts of varying levels of difficulties. This new era, as noted by Hiebert and Pearson (2014), was not a comfortable context for traditional readability formulas and their narrowed focus on surface-level features as sole predictors of discourse comprehension.

The validity of these formulas is commonly established by correlating readers' performance on a reading comprehension test with the readability score computed by the formula (Cunningham & Mesmer, 2014). However, the use of a comprehension test as an independent measure or indicator of text difficulty is problematic since it is hard to know whether text complexity is due to difficulty in the text or the task employed (Bailin & Grafstein, 2016). As an alternative method, the "cloze procedure" (W. L. Taylor, 1953) is often used. The cloze test involves deleting parts of texts at random and asking readers to fill in the missing words. The readability of a text is assessed by the extent to which readers can provide the deleted words. Readers need to identify patterns within the test using their background knowledge of the content to guess the missed words. Cloze tests do not assume that text difficulty is due to complex words and long sentences but rather leave open the possibility that the reader may know common polysyllabic words, while at the same time he or she may not be familiar with uncommon monosyllabic words (Bailin & Grafstein, 2016). However, the use of the cloze test is controversial (Cunningham & Mesmer, 2014). For instance, Shanahan, Kamil, and Tobin (1982) have questioned whether it is a valid test of comprehension beyond the sentence level.

Notwithstanding the criterion variable on which they were developed and validated, traditional formulas have been criticized for their simplistic and shallow approach which ignores key textual proprieties of text known to impact deep-level processing of text, such as cohesion, syntactic ambiguity, rhetorical organization, and propositional density (see, e.g., Crossley, Skalicky, Dascalu, McNamara, & Kyle, 2017; Davison & Kantor, 1982). Sawyer (1991) argued that the early readability formulas were "misleading and overly simplistic" (p. 309). Similarly, Coupland (as cited in Klare, 1984) noted that "the simplicity of … readability formulas … does not seem compatible with the extreme complexity of what is being assessed" (p. 15). Despite they have been on the receiving end of much criticism, readability formulas are still in use for determining text readability, mainly due to their ease of use and intuitive interpretation.

Nelson, Perfetti, Liben, and Liben (2012) assessed the capabilities of seven text difficulty metrics, including Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004), Lexile A. J. Stenner, Horabin, Smith, and Smith (1988), and SourceRater (Sheehan, Kostin, & Futagi, 2010), in providing more accurate prediction of difficulty in external, independent graded texts. While many of these tools rely on some variant of measures of word difficulty (length, frequency) and sentence complexity, other tools use indices related to text cohesion, e.g., Coh-Metrix. The results indicated that text difficulty tools "that included the broader range of linguistic and text measures produced higher correlations than the measures that [relied exclusively on] word difficulty and sentence length measures" (Nelson et al., 2012, p. 4). Nelson et al. (2012) also found that the seven text difficulty metrics performed differently on informational texts and narrative texts, perhaps due to the impact of text genre on text difficulty estimation.

#### 2.1.1.2 Cognitively Inspired Methods

As text readability researchers began to know more about the processes involved in text comprehension, it has become evident that shallow textual features are no longer adequate to explain an intricate cognitive phenomenon as comprehension (Crossley, Skalicky, & Dascalu, 2019). Accordingly, new alternative approaches have been developed in accordance with cognitive theories of text comprehension that explain how human store and retrieve information in long-term memory, such as schema (e.g., R. C. Anderson &

Pichert, 1978), prototype (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976), connectionist (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982) theories. Cognitively inspired software of text complexity, e.g., Coh-Metrix (Graesser et al., 2004; McNamara, Graesser, McCarthy, & Cai, 2014) and WordNet (Miller, 1998; Miller, Beckwith, Fellbaum, Gross, & Miller, 1990), go beyond the analysis of surface-level text features considered by traditional formulas to include analyses of discourse properties, such as discourse cohesion and coherence (Graesser et al., 2004).

Based on text comprehension models (e.g., Kintsch & Van Dijk, 1978), researchers hypothesize that text complexity is more related to textual cohesion and how different elements of the text interact with each other (Britton & Gülgöz, 1991; Gernsbacher, 1990; Kintsch, 1988; McNamara & Kintsch, 1996a). Cohesion, being a defining feature of text (Halliday & Hasan, 1976), is a central indicator of text readability in these approaches. Less cohesive texts are assumed to be more difficult as readers need to invoke more mental resources to fill in gaps in the text (Zwaan & Radvansky, 1998). In a highly cohesive text, propositional or argument overlaps among successive sentences (Kintsch & Van Dijk, 1978). Put it differently, when there is a high propositional overlap in a text, inferences become more explicit, and the reader does not need prior knowledge about the topic to fill in gaps in the text (Benjamin, 2012).

Two common measures of text cohesion, referential and causal cohesion, are often considered when measuring text cohesion (Sheehan et al., 2010). Referential cohesion refers to the degree to which words, phrases, or concepts are repeated across a text whereas causal cohesion refers to the degree to which causal relationships are explicitly stated in a text, for example, using connectives such as because, therefore, and consequently (McNamara et al., 2014). Cognitively inspired tools, e.g. Coh-Metrix (McNamara et al., 2014), often analyze hundreds of linguistic features assumed to affect lexical familiarity, syntactic difficulty, and text cohesion. For example, Coh-Metrix software computes over 200 linguistic indices of text complexity. The software also conducts Latent Semantic Analysis (LSA) to analyze the semantic relatedness between texts or among different units or segments of a text (McNamara et al., 2014). LSA is discussed in greater detail by Foltz, Kintsch, and Landauer (1998) and Landauer, Foltz, and Laham (1998).

Readability researchers have also utilized statistical language models in their approach to assessing test readability. Formally, a language model can be defined as predicting a probability distribution of words from the fixed size vocabulary $V$, for word $w_{t+1}$, given the historical sequence $w_{1:t} = [w_1, ..., w_t]$. One approach is to train separate n-gram SLMs on a large corpus of texts for each readability class (e.g., Collins-Thompson & Callan,

2005; Si & Callan, 2001). Besides requiring a large annotated corpus, this approach did not improve accuracy. A more successful approach is to use statistical scores derived from language models (log-likelihood and perplexity scores) as features along with others as input for another classifier (e.g., Petersen & Ostendorf, 2009; Schwarm & Ostendorf, 2005; Xia, Kochmar, Briscoe, & Building, 2016). As text complexity features are becoming increasingly fine-grained and technically complicated, it may become bewildering, to say the least, for novice researchers and teachers to interpret and make use of the results generated by machine learning models. Furthermore, the detailed results that these tools generate may also suggest no further analyses are required. However, as warranted by the original creators of Co-Metrix (McNamara et al., 2014), any predictions based on Coh-Metrix values should be interpreted with careful consideration of the multiple real-world factors that surrounded the text, including factors related to the reader and task, and how these factors potentially interact with text features.

#### 2.1.1.3 Machine Learning Approaches

The recent advances in computational linguistics and the availability of large text corpora have allowed readability researchers to systematically examine hundreds of linguistic features using more sophisticated machine learning algorithms (Benjamin, 2012; Collins-Thompson, 2014). Early works using this approach begin in the early-mid 2000s (Collins-Thompson, 2014). The use of more advanced machine learning algorithms has proved to be useful and generally leads to a more accurate assessment of text difficulty than traditional readability formulas. While some machine learning approaches to readability assessment appear similar to traditional readability formulas in the sense that they rely on linear regression, they nonetheless incorporate a wider range of features that tap into different dimensions of text complexity (Collins-Thompson, 2014). Consequently, they are more ecologically valid than traditional readability formulas which exclusively depend on surface-level features for their estimation of text difficulty. Additionally, unlike traditional readability formulas, machine learning algorithms can deal with 'noisy data' and ill-formed sentences often found in web texts (Collins-Thompson & Callan, 2005; L. Feng, Elhadad, & Huenerfauth, 2009; Peterson, 2006).

As a supervised machine learning problem, text readability can be described as a task of learning a mathematical function that maps an input (a matrix of complexity features) to an output (a vector of readability score or level) based on examples of input-output pairs—or formally written as $\{(x_1, y_1), (x_2, y_2), ...., (x_n, y_n)\}$. Depending on what task or

application it would serve, a readability assessment tool or system can be configured for the task of classifying, predicting, or ranking texts based on their levels of difficulty.

The flexibility of machine learning algorithms along with the advances in computational linguistics has enabled readability researchers to further explore new aspects of text readability and devise features that are theoretically and empirically supported by research in text processing and comprehension (Benjamin, 2012; Collins-Thompson, 2014). Recently, readability researchers have experimented with ensemble learning and hierarchical machine learning approaches (e.g., Balyan, McCarthy, & McNamara, 2020). The results showed these approaches improve the accuracy of text difficulty prediction.

While machine learning approaches often yield more accurate prediction of text readability, they, however, have some caveats. Firstly, for a more accurate prediction of text readability, a machine learning algorithm needs to be trained on a large corpus of human-labeled texts, which is very scarce. Constructing a gold-standard corpus is a labor-intensive process. To mitigate this obstacle, readability researchers have utilized crowdsourcing services (e.g., Amazon Mechanical Turk) to annotate their text corpora in a more efficient way (Crossley et al., 2017; De Clercq et al., 2014). However, a proper annotation scheme should be in place for accurate and consistent labeling of text. In a recent study, Zampieri, Malmasi, Paetzold, and Specia (2017) evaluated the performance of different computational systems on an annotated dataset and found that these systems predict text complexity poorly. Zampieri et al. (2017) attributed this poor performance to how inaccurately human annotation was performed.

#### 2.1.1.4 Deep Learning Approaches

The use of deep neural networks in readability assessment is relatively new (e.g., Deutsch et al., 2020; Filighera et al., 2019; Martinc et al., 2019; Nadeem & Ostendorf, 2018). A key motivation for the use of deep neural networks is their capability of self-learning, with minimum involvement or 'supervision' of researchers. That is, given a large corpus of data and target labels, artificial neural networks learn underlying patterns or features of texts that make them hard. However, to learn useful features, a deep neural network requires a large collection of texts from the same domain. Martinc et al. (2019) investigated the utility of different neural architectures with automatized feature generation and compared their performances with that of classical machine learning algorithms. Particularly,

the researchers found that a pre-trained transformer model achieved state-of-the-art performance on the WeeBit corpus while a hierarchical attention network (HAN) achieved state-of-the-art performance on the Newsela corpus (Xia et al., 2016).

Recently, Deutsch et al. (2020) explored whether augmenting deep neural networks with linguistically motivated features leads to better performance. Specifically, the researchers fused the output of the neural language model with hand-crafted linguistic features which were used as input to machine learning algorithms such as SVM. The researchers found their proposed approach did not seem to lead to any better prediction of text readability than what was offered by traditional machine learning approaches.

However, given their mathematical working, deep learning approaches are often conceived as 'black box' which can be impenetrable to human comprehension. This may not be a problem if the primary goal of text difficulty analysis is to accurately predict the difficulty of new texts—which is often the case. As research in deep learning is rapidly developing, we anticipate the development of more readability assessment tools based on deep neural networks in the coming few years. But as of now, the usefulness of this approach remains to be determined by the research community.

### 2.1.2 Second Language Text Difficulty

Insofar, I reviewed key major approaches and methods for assessing text readability for native readers. As I previously mentioned at the beginning of this chapter, there is an extensive body of literature on L1 text readability and a diverse set of tools for assessing and selecting texts for L1 readers. However, to date, little research has been carried out on L2 text complexity and difficulty. In what follows, I review the small, yet growing, body of research on assessing the difficulty of written and spoken texts for English as a Second Language (ESL) learners (e.g., Brown, 1998; Fulcher, 1997; Vajjala & Meurers, 2012; Xia et al., 2016; S. Yoon, Cho, & Napolitano, 2016).

#### 2.1.2.1 L2 Written Text Difficulty

It seems to be beyond serious dispute that what characteristics of text make it difficult for L1 and L2 readers are not the same. Grammatical and lexical knowledge of the second language is too easy to point out. While L1 readers usually have a good understanding of grammatical systems of their first language even before commencing proper schooling,

L2 readers, on the other hand, are still in the process of developing their knowledge of their new language and its grammar. The same can also be said for lexical knowledge. Due to the lack of purposefully developed formulas for L2 readers, some L2 readability scholars have explored the possibility of using L1 readability formulas to select texts for L2 readers. For example, in his pioneering study, Brown (1998) compared the mean cloze scores of 2,300 Japanese EFL learners on English reading books with the scores predicted by six L1 readability formulas, including Flesch, Fry, and Flesch-Kincaid. The resulting correlations ranged from .48 to .55, leading Brown to conclude that "first language readability indices are not very highly related to EFL difficulty" (Brown, 1998, p. 27). Accordingly, Brown developed a new EFL formula using variables (e.g., the average number of syllables per sentence and the percentage of words in the text of more than seven letters) that were highly predictive of L2 text difficulty. The new formula performed well and accounted for more variance in EFL text difficulty than did the traditional readability formulas. Similarly, Greenfield, Jerry (2004) found that L1 readability formulas do not work well on predicting EFL text complexity.

Crossley, Greenfield, and McNamara (2008) developed an L2 text complexity formula, called Coh-Metrix L2 reading index, to measure English text readability for L2 readers. The researchers found that three linguistic variables related to cognitive reading processes—lexical coreferentiality, sentence similarity, and word frequency—yield more accurate prediction of L2 text complexity than the simple, surface-level variables used in traditional L2 formulas. Despite being developed using a small set of passages, Crossley et al. (2008) argued that their proposed formula is better suited to predict the readability of L2 texts because it relies on more reliable psycholinguistic factors in reading comprehension such as decoding, syntactic parsing, and meaning construction. Later, Crossley, Allen, and McNamara (2011) demonstrated that their L2 formula, Coh-Metrix L2 Reading Index, outperformed traditional readability formulas on a large corpus of texts intuitively simplified for language learners. Their findings suggested that the psycholinguistic variables adopted in the index seem to closely align to the intuitive text processing employed by the authors when simplifying their texts. However, two limitations can be leveled against the L2 Reading Index. First, the formula was developed using a fairly small homogeneous text corpus, 31 academic articles. Thereby, the formula validity and its generalizability to other types of text are questionable. The second limitation is that in their study, Crossley et al. (2008) have used the cloze test, which has become known to be a less valid indicator of text comprehension, to generate the difficulty of their texts.

Recently, researchers have developed L2 text difficulty tools based on machine learning

algorithms (e.g., Sung, Lin, Dyson, Chang, & Chen, 2015; Thomas & Cédrick, 2012; Xia et al., 2016). For example, Schwarm and Ostendorf (2005) developed a readability assessment model using Support Vector Machines (SVMs). In training their classifiers, they used several traditional readability features and Language Models (LM) features (e.g., word unigrams, bigrams, and trigrams) to classify articles as belonging to one of five grades. Articles in their corpus were collected from four different sources, e.g., *Weekly Reader, Britannica Elementary Encyclopedia, CNN News Stories, Encyclopedia Britannica.* The model performance was evaluated using different metrics: detection error trade-off (DET) curves, precision, and recall. The highest accuracy the classifiers achieved was a precision of 75% and a recall of 87%. In a subsequent study, Petersen and Ostendorf (2009) extended their features set and developed SVM classifiers that incorporate features based on an automatic parser, n-gram LMs (based on perplexity scores of the models), as well as traditional readability features. The evaluation results showed that the SVM classifier outperformed both Flesch-Kincaid and Lexile metrics.

Sung et al. (2015) developed a readability assessment system for classifying or leveling Chinese texts according to the Common European Framework of Reference (CEFR) levels. The authors trained a Support Vector Machine (SVM) model on 1,578 CFL texts judged by CFL teachers as belonging to one of the CEFR levels. A set of 30 CFL readability features were used on the training samples and sorted based on their importance weights using F-scores. The empirical evaluation of the model revealed high accuracy (average exact-level 74.97% and adjacent-level, 99.62%) for predicting expert classification of a text. Data-driven approaches to text readability assessment are more accurate than other approaches, yet they require large-sized, properly annotated datasets for accurate evaluations (Xia et al., 2016).

Vajjala and Meurers (2012) adopted measures from Second Language Acquisition (SLA) literature and used them along with traditional readability measures to build their classification models. Using different classification algorithms, such as Decision Trees, Multilayer Perceptron (MLP), Support Vector Machines, and Logistic Regression, Vajjala and Meurers (2012) found MLP to be the best classifier which achieved a classification accuracy of 93.3%. The results demonstrate that SLA features help in determining text complexity. Vajjala and Meurers (2014) trained classifiers based on a set of features derived from research on text complexity in SLA and psycholinguistics. The classifiers achieved a classification accuracy of 95.9% on a corpus of BBC TV subtitles of age-specific TV programs.

Kotani and Yoshimi (2017) proposed a method that incorporates learner and linguistic features to explain text readability for language learners. The learner features include features such as learners' English learning experience, experiences of visiting English countries and places, frequency of reading, and their scores on reading comprehension tests. The results demonstrate that the combination of the learner factors and acoustic features improve the measurement of EFL text readability, though not all learner features contributed significantly. For example, the learners' language learning experiences did not predict how learners judge the readability of EFL texts.

Vajjala and Lucic (2018) trained classifiers on generic text classification features (e.g., n-grams and dependency relations) and other features commonly used in ARS research. The results demonstrate that the character n-gram classifier achieved a classification accuracy of 77% and it is less than one point of the highest classification accuracy (78.13%) achieved when all features are included.

The field of L2 text difficulty assessment is still in its infancy and given the importance of comprehensible input for learning acquisition, there is a need for further research into the impact of text difficulty on language learners (Xia et al., 2016). Specifically, the field needs automated tools tailored to second language learners. Automated text difficulty tools have many applications in second language teaching and learning. They can become very handy for language teachers, material writers, and textbook editors who need to select texts or write new content for language learners. Text simplification is another application wherein text complexity tools are used to modify texts; that is replacing complex features, e.g., complex sentence structure, and rare and sophisticated words, with simple ones (Crossley, Allen, & McNamara, 2011). Aside from text simplification, text complexity measures are used in analyzing language learners' speaking and writing (e.g., Révész, Kourtali, & Mazgutova, 2017; H. J. Yoon, 2017) and in developing automatic scoring systems for assessing L2 speaking (e.g., L. Chen et al., 2018) and writing performance (e.g., Burstein, Tetreault, & Madnani, 2013). The construction and public release of L2 text corpora, e.g., OneStopEnglish (Vajjala & Lucic, 2018) and Cambridge English Exam data (Xia et al., 2016), would make developing and validating new models eminently feasible.

#### 2.1.2.2 L2 Spoken Text Difficulty

Spoken text complexity, also called listenability, is concerned with analyzing spoken texts to identify features of spoken language that impede speech comprehension. Thus far, only a handful of studies have attempted to measure L2 spoken text complexity (e.g., Kotani

& Yoshimi, 2017; Vajjala & Meurers, 2014; S. Yoon et al., 2016). Understanding what may cause difficulty in comprehending spoken language requires researchers to pay special attention to acoustic and phonological aspects of spoken language, such as pitch, stress, speech, articulation rate, and so forth. Spoken texts, especially naturalistic ones, are replete with speech disfluencies, overlapping speech, ungrammatical utterances, semantically vague expressions, and slangs (H. H. Clark, 1996). Furthermore, the two types of text have different structural characteristics (see e.g., Biber, 1986, 1988; Flowerdew & Miller, 2005). Spoken text is loosely structured and fragmented whereas written text is more densely structured. While useful, researchers also need to go beyond spoken texts and understand how language learners perceive and deal with varieties in spoken texts. Research in second language listening comprehension has identified some aspects of spoken language that are more challenging for language learners (see Bloomfield et al., 2010, for a comprehensive review).

Previous studies have investigated different acoustics and phonological features along with other features of spoken texts in developing a measure of spoken text complexity for language learners (e.g., Kotani & Yoshimi, 2016; S. Yoon et al., 2016). For example, S. Yoon et al. (2016) found that combining acoustic features, e.g., speaking rate, number of silences per word, and variations in vowel duration, with lexical and grammatical features lead to a better estimate of spoken text difficulty. Similarly, Kotani and Yoshimi (2016) developed a decision-tree classifier using a set of acoustic and linguistic features including speech rate, elision, and phonological reduction and contraction. They found that speech rate is one of the most discriminative features of listenability.

## 2.2 Principles of Automated Text Difficulty Assessment

While they have made great strides over the years, computational approaches to text difficulty assessment have received some criticisms concerning their reliability, validity, and whether they actually provide informative feedback on text difficulty (e.g., Bailin & Grafstein, 2001; Fisher, Frey, & Lapp, 2016; Redish, 2000). I argue that much of these concerns stem from the mistaken assumption that computational text difficulty assessment tools can tap into and objectively measure all facets of text comprehension. Virtually all comprehension theories assert that text understanding is a multifaceted phenomenon, best conceived as a function of the integration between a multitude of factors germane

to texts, readers, and tasks. It is therefore unreasonable to assume that text difficulty assessment models, irrespective of how complicated or advanced they are and how many variables they include, can fully assess as a complex phenomenon as comprehension and for a group of different individuals.

Equally disturbing is the use of text difficulty metrics and tools for assessing a wide variety of texts and for readers and tasks beyond those for which the formulas were originally created (Davison & Kantor, 1982; Greenfield, 1999). When a text readability metric is developed and becomes easily accessible, (e.g., integrated into Microsoft Word or Grammarly), it is easy to imagine that it is used by the general public without any consideration for what type of texts and for whom they are designed. Another false assumption is that rather than reflecting readability text difficulty metrics *define* readability (Davison & Kantor, 1982). This assumption had led many publishers and writers, under industry pressure, to use readability formulas as a guide to compose or modify existing texts to meet a certain readability level. Efforts to simplify or "dumb down" texts, for example, by shortening their sentences and replacing long and complex words with shorter and simpler ones, often inadvertently lead to texts that are more difficult to comprehend (Davison & Kantor, 1982; Goldman & Rakestraw Jr, 2000). This is mainly because the revised texts often require readers to make more inferences and compromise for the lack of explicit references (McNamara & Kintsch, 1996b).

In this section, key issues in developing text difficulty assessment are raised to illuminate their potential impact on developing more robust measures and hence more accurate feedback about text difficulty. While I discuss these principles with L2 language learners in mind, they are generic enough to be used for developing text difficulty tools for first language speakers as well.

### 2.2.1 A Definition of Text Difficulty

Text difficulty is one of those notions that are presumed to be known by everyone, and hence many researchers feel no obligation to define it. Notwithstanding decades of research, the notion of difficulty remains elusive and is often used interchangeably with terms such as complexity and comprehensibility (Amendum, Conradi, & Hiebert, 2017). One obvious reason for this nomenclature is the inherent ambiguity of the term itself.

Neglecting for a moment the notion of difficulty, there are at least two meanings of complexity in most dictionaries as 1) "composed of two or more parts" or/and 2) "hard to

separate, analyze, or solve" (Merriam-Webster, 2021b). The two meanings of complexity are bewildering already, or maybe even worse they lead to terminological circulation; for instance, a linguistic feature is difficult because it is complex, or a linguistic feature is difficult because it is complex.

Noting the importance of differentiating the two constructs, Mesmer, Cunningham, and Hiebert (2012) argued that text complexity and difficulty are two conceptually separable constructs. On the one hand, text complexity refers to the characteristics of individual texts, e.g., sentence length, that can be analyzed and manipulated irrespective of the readers/listeners or task. On the other hand, text difficulty refers to how easy or difficult a text is for individual readers. Put it differently, the complexity of text or a feature always implies an independent or predictor variable and the difficulty of a text or feature implies a dependent or criterion variable, hence "treating the terms text complexity and text difficulty as synonymous conflates causes with effects" (Mesmer et al., 2012). However, Sheehan, Kostin, Napolitano, and Flor (2014) noted that while researchers often collect different types of evidence for text complexity and difficulty, they are not fundamentally measuring two different constructs, but rather "a common text characteristic: the ease or difficulty of forming a coherent mental representation of the information, argument, or story presented in the text" (185).

Terminological confusion aside, researchers have ascribed text readability or the lack thereof to numerous factors. Consider the following definitions of text difficulty/readability:

- "In general, we have used the words legibility and readability interchangeability to mean 'ease and speed of reading printed material at a natural reading distance'" (Paterson & Tinker, 1940, p. 9)
- "The total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting" (Dale & Chall, 1949, p. 23)
- "the ease of understanding or comprehension due to the style of writing." (Klare et al., 1963, p. 11)
- "the degree to which a given class of people find certain reading matter compelling and comprehensible."
- "the ease of reading created by the choice of content, style, design, and organization that fit the prior knowledge, reading skill, interest, and motivation of the audience" (DuBay, 2007, p. 6)

- "a subjective judgment of how easily a reader can extract the information the writer or speaker intended to convey" (Kate et al., 2010, p.548)
- "inherent difficulty of reading and comprehending a text combined with consideration of reader and task variables" (NGACB & CCSSO, 2010, p. 43)

The definitions laid out above clearly show that there is no consensus among readability scholars on what makes a text to be more or less difficult to read. It is therefore crucial that text difficulty researchers define text difficulty because it is through this definition the whole enterprise of assessing text difficulty will be operationalized and against which it will be validated.

### 2.2.2 A Proper Measurement of Text Difficulty

Once a definition of text difficulty is adopted or provided, researchers need to use a proper way to measure it. Klare (1984) rightfully noted that computational text difficulty tools "can be no more accurate than the criteria on which they are based" (p. 701–702). Cunningham and Mesmer (2014) observed that text researchers generally pay more attention to text predictors in a tool than the criterion variable that was used to develop the tool.

Over the decades, text difficulty researchers have employed different approaches for establishing a criterion variable to be used in the development of text difficulty assessment metrics. For example, text researchers have used comprehension tests, cloze procedures, reading speed, other formulas, and subjective ratings made by students, teachers, or text publishers. However, what approach best reflects text difficulty is a matter of debate.

Though they are not often explicitly stated, two primary factors seem to be influencing researchers' decision on what approach they adopt for determining a criterion variable for text difficulty. First and foremost, how text researchers define the construct of text difficulty reflects what approach they ended up choosing. For some researchers, text difficulty and comprehension are essentially the same constructs; thereby a reader's or listener's performance on a comprehension test should reflect the difficulty or ease in text. For other researchers, text difficulty is merely subjective and highly dependent on the reader/listener's perception of the text. Still, others believe that readability has to do with ease of processing and how comfortable a reader is when reading a text (Crossley, Skalicky, & Dascalu, 2019; Dale & Chall, 1949; Newbold & Gillam, 2010; Richard, Platt, & Platt, 1992). As such, some measures of reading speed should be used to evaluate text difficulty (Crossley, Skalicky, & Dascalu, 2019).

The second factor is convenience or practicality. In some situations, obtaining a more direct and objective measure of text difficulty is not feasible. Therefore, text difficulty researchers opt for more practical approaches, e.g., subjective rating and crowdsourcing, especially for developing machine learning models of text difficulty which require a larger volume of labeled texts for training and validation.

Historically, teacher subjective judgment of text difficulty was the most commonly used approach for determining and selecting texts (Cunningham & Mesmer, 2014; Pearson & Hiebert, 2014), but there has been a gradual shift towards reader performance measures (Cunningham & Mesmer, 2014). One such popular measure is cloze tests (W. L. Taylor, 1953). W. L. Taylor proposed the cloze technique as an alternative approach to readability formulas. Soon afterward, the approach became a popular means for assessing text comprehension, replacing the usual multiple-choice test in the measurement of readability. For example, it was employed as a criterion variable in the development of the Bormuth formula (Bormuth, 1968), the Dale-Chall formula (Chall & Dale, 1995), and the Coh-Metrix L2 Readability Index (Crossley et al., 2008).

In its most common form, the cloze procedure involves the deletion of words randomly or fixedly. For example, every tenth word in a passage is deleted and text difficulty is then determined by calculating the probability of guessing the correct fill-in. The deletion takes place regardless of the importance of the word or its grammatical function (Bailin & Grafstein, 2016). The central assumption behind the cloze procedure is that filling in the missed words is more accessible in less complex text than it is in more complex texts. The cloze technique is thought to be a reflecting indicator of text difficulty as experienced by the reader as successful fill-ins depend heavily on the reader's ability to understand the text, extract and integrate key ideas, and evaluate information across sentences and paragraph boundaries. Modified cloze techniques have also been used in the development of text complexity tools . For example, J. A. Stenner (1996) used a modified cloze technique in the development of the Lexile Framework wherein students in grades 2 through 12 were provided with a sentence with one word blanked out and four possible option words for each text they read. Students' responses then were subject to Rasch analyses which helped to determine the difficulty level of the texts.

Several studies have evaluated the cloze approach (McKenna & Layton, 1990; Shanahan et al., 1982). Generally, previous research has challenged the applicability of the cloze procedure as a global measure of text comprehension. For example, Shanahan et al. (1982) compared readers performance on cloze passages under four conditions: (1) intact passages, (2) passages with scrambled sentence, (3) intermingled passages with embedded

sentences from other passage, and (4) eclectic passages that are composed of unrelated sentences. Shanahan et al. (1982) found no performance differences among readers under all conditions and concluded that the cloze test does not seem to provide useful information about readers' ability to use information across sentence boundaries (this ability is commonly referred to as intersentential comprehension). Other studies have also arrived to the same conclusion (e.g., Kintsch & Yarbrough, 1982).

Cunningham and Mesmer (2014) argued that direct approaches such as comprehension and cloze tests should be the gold standard approach for determining text difficulty. However, comprehension data is expensive to obtain especially on a large scale. Besides issues of practicality and more importantly, comprehension tests may not serve as valid measures of deep levels of comprehension (Crossley et al., 2017; Magliano, Millis, Ozuru, & McNamara, 2007).

Aside from comprehension measures, a human judgment of text difficulty is still commonly used to assign a grade level or a numerical value indicating the reading or listening difficulty of texts. This approach is becoming increasingly popular; for example, it is commonly used by textbook publishers and in the development of some commercial text difficulty measures, e.g., SourceRater (Sheehan et al., 2010). Subjective evaluation of texts can be made by language experts, teachers, or students.

Although subjective evaluation remains a labor-intensive process for large-scale projects, it is nonetheless more practical than comprehension tests. Recently, researchers have explored the use of crowdsourcing as a way to reduce the cost of collecting human judgments of text difficulty (e.g., Crossley et al., 2017; De Clercq et al., 2014). Another approach was described in Heilman, Collins-Thompson, Callan, and Eskenazi (2007) in which a readability model was developed based on a collection of texts compiled from websites that create reading materials for students at different grade levels.

For developing L2 text difficulty tools and metrics, the Common European Framework of Reference for Languages (CEFR) (of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, 2001) is often used (Pilán, 2018). The CEFR provides general guidelines, in form of "can-do" statements, that can help to define L2 language skills and competence into six levels of language abilities: break-through (A1) and waystage (A2), threshold (B1), and vantage (B2), and effective operational proficiency (C1) and mastery (C2). The CEFR has been and still is widely used in the development of many language teaching and learning materials textbooks and several international English proficiency tests, such as TOEFL and TOEIC have been mapped onto the CEFR.

Using corpora annotated per the CFER levels, researchers have developed readability systems for classifying reading materials in languages such as English (T. Arnold, Ballier, Gaillat, & Lissòn, 2018; Xia et al., 2016), Chinese (Sung et al., 2015), French (Thomas & Cédrick, 2012), Portuguese (Curto, Mamede, & Baptista, 2015), Russian (Karpov, Baranova, & Vitugin, 2014), and Swedish (Pilán, 2018).

In most cases, researchers have built their CEFR corpus by compiling already annotated coursebooks and exams (Xia et al., 2016), and very few have built their own corpus using authentic texts written primarily for L1 readers, but then were assigned CEFR level by a committee of teachers (Sung et al., 2015). The availability of gold-standard CFER-based corpora is very limited, mainly due to copy-right issues and the annotation cost that is often involved in building a large corpus suitable for machine learning experimentation (Xia et al., 2016). Moreover, the CFER descriptors (can-do statements) have been criticized for being opaque and subjective, with different educators, editors, textbook publishers interpreting them differently (Sung et al., 2015). Additionally, educators need proper training for effectively using the framework, for example, in selecting texts for a target language level (Alderson, 2007).

Following the above discussions, it should be clear that there is no single approach for obtaining a reference score for text difficulty, but many. Unfortunately, there is little information available on which criterion variable seems to better reflect text difficulty. Indeed, a century worth of research has preliminary focused on textual elements associated with text difficulty, and very little research was devoted to evaluating the criterion variables on which text difficulty tools are calibrated (e.g., Cunningham & Mesmer, 2014).

### 2.2.3 An Inclusive Treatment of Text Complexity

As discussed earlier, traditional readability formulas were criticized for affording an overly simplistic approach to the estimation of text readability (Bailin & Grafstein, 2016; Crossley et al., 2017). Several readability scholars contended that the mere reliance on sentence length and word frequency does not seem compatible with the intricacy and complexity of what is being measured (e.g., Sawyer, 1991). As such, many classical readability tools suffer from an "inadequate construct coverage" (Sheehan et al., 2010, p. 1).

To provide a more comprehensive treatment of text complexity, readability researchers have attempted to align their conceptualization of text difficulty with well-established

theories of text comprehension. This has allowed them to explore new aspects and dimensions of difficulty that go beyond the conventional dimensions of semanticity and syntactic complexity, including, for example, issues pertaining to text narrativity and cohesion (Gernsbacher, 1990).

As a result of incorporating these new dimensions of text complexity into the assessment of text difficulty, text difficulty scholars have not only strengthened the validity of the construct but the use of these also proven empirically beneficial. Indeed, a substantial body of empirical studies has confirmed the effectiveness of using complexity measures that tap into multiple dimensions of text in assessing text readability (e.g., Sheehan, 2017).

Having said that, data-driven approaches to text difficulty assessment can also be beneficial. In fact, great insights can be generated from empirical analyses performed on large datasets. But, to explain these insights, theories are necessary.

### 2.2.4 Controlling for Text Variations

While it is probably self-evident that different types of text (news articles vs. scientific articles) impose different demand on readers, many text difficulty assessment tools, e.g., the Flesch-Kincaid Grade Level formula (Kincaid et al., 1975); the Lexile Framework (A. J. Stenner et al., 1988), the Coh-Metrix L2 Readability Index (Crossley et al., 2008), and the Degrees of Reading Power (Carver, 1985), do not account for genre differences. This may explain why text difficulty measures often yield inconsistent estimates when used with texts belonging to different genres (Sheehan, 2017).

It is conceivable that text genres are not equally complex; a narrative text, for example, is less difficult and more pleasing to read than a scientific text. Cognitive researchers have found empirical evidence suggesting a link between text genres and a reader's ability to understand a text (Cervetti, Bravo, Hiebert, Pearson, & Jaynes, 2009). Text genre also influences the frequency of core vocabulary usage (S. S. Lee, 2001), word reading speed (Zabrucky & Moore, 1999), the number of prepositions recalled (Graesser, Hauft-Smith, Cohen, & Pyles, 1980), and the types of inferences generated (Best, Floyd, & Mcnamara, 2004), and the types of prior knowledge accessed during inference generation (Best et al., 2004). Additional support comes from large corpus analysis studies which demonstrated syntactical and lexical differences across text genres. For example, D. Y. W. Lee (2001) conducted an analysis of core vocabulary usage variation within a corpus of literary and informational texts compiled from the British National Corpus, an over one million word

corpus. Core vocabulary was operationalized as the most 2000 common words that are listed in the dictionary definitions presented in the Longman Dictionary of Contemporary English. The analyses revealed that literary texts had a higher percentage of core vocabulary usage than informational texts.

Both linguistic and cognitive indicators of genre effect on text difficulty provide a strong argument for developing genre-specific models that are optimized for predicting difficulty in a given type of text. When they are not developed for a specific type of text, text difficulty systems are more prone to *genre bias*; that is "the tendency of some measures to overpredict the complexity levels of informational texts, while simultaneously under predicting the complexity levels of literary texts" (Sheehan et al., 2014, p. 198). Sheehan et al. (2010) provided empirical evidence suggesting that many linguistics complexity indices used in common readability systems function differently across text genres.

Addressing genre bias, Sheehan, Flor, and Napolitano (2013) proposed a two-stage approach for generating unbiased estimates of text complexity. At the first stage, texts are classified as belonging to one of three possible genres: literary, informational, or mixed texts. Text genre classification can be done by either human annotators or an automated genre classifier. Next, at stage two, three optimized prediction models are used to generate a complexity score for each text. Each model is trained on a set of features extracted from texts belonging to a correspondent text genre.

Dell'Orletta, Venturi, and Montemagni (2012) demonstrated the shortcomings of general-purpose readability assessment tools in reliably predicting the readability of texts belonging to different genres and they suggested the use of genre-specific models. In a recent study, Toyama, Hiebert, and Pearson (2017) found a clear variability in scores projected by analytical tools despite being applied to the same texts. It is not surprising because many text complexity tools are developed and validated on a collection of texts belonging to a particular text genre. Therefore, they should not be used to evaluate the complexity of texts belonging to a different genre. Most text readability metrics are developed and tested on datasets composed entirely of informational texts, e.g., news articles. Many recent text complexity studies have used datasets compiled from informational texts only, e.g., Wall Street Journal articles (Pitler & Nenkova, 2008), Encyclopedia Britannica articles (Schwarm & Ostendorf, 2005) and Weekly Reader articles (Vajjala & Meurers, 2012). Therefore, researchers have recently explored the benefits of developing readability measures tailored for a particular genre (e.g., Ma & Fosler-lussier, 2012; Nadeem & Ostendorf, 2018; Sheehan et al., 2013) or reader group (Ozasa, Weir, & Fukui, 2014). In the last few years, several genre-specific models have been developed (Dell'Orletta et al., 2012;

Nadeem & Ostendorf, 2018). For example, Nadeem and Ostendorf (2018) developed text difficulty models for estimating readability in science texts using Recurrent Neuron Networks (RNNs).

Besides text genre bias, researchers may also need to control for drastic variations in text length. This is because some text complexity measures are more sensitive to text length and hence their use may impact the performance of automated text difficulty tools. For instance, indices of lexical diversity measures, such as Type-Token Ratio (TTR), are more susceptible to text size (see Chapter 3 for more discussion). It is therefore recommended that text difficulty researchers use more robust indices.

### 2.2.5 Controlling for Learner Variables

While text complexity plays a critical role in the determination of text ease or difficulty, learner factors and how they interact with texts cannot be ignored (NGACB & CCSSO, 2010). In fact, a text is conceived as easy or difficult will depend upon, inter alia, the learner's language ability and skill, experience, familiarity with any terminology, and purpose of reading or listening. In L2 research, it has been widely acknowledged that learner factors significantly impact language learning and acquisition (see Dörnyei, 2014, for a comprehensive account of learner factors)

Methodologically, it is very challenging, if not impossible, to examine all text and learners' factors and their complex interactions in a single study. A more reasonable approach, therefore, is to concentrate on understanding text complexity first and once we know what makes a text difficult, we can then explore how language learners, with different abilities and traits, contend with difficult and easy texts, and under what task conditions and within what learning contexts text comprehension is enhanced or hampered. However, in the interim, researchers need to control for learner factors and limit their interference with learners' perception of text difficulty. One possible way is to develop a text difficulty tool for a specific group of readers or listeners, who share some common characteristics, e.g., school grade, age, cultural background, gender, and reading and licensing ability.

### 2.2.6 Validation

Once a computational model of text difficulty is developed, it needs to be validated; that is to test if the developed model or system is yielding consistent results when applied on unseen data, i.e, out-of-sample data that were not used to fit the model. There are several validation approaches in text difficulty assessment literature and in the field of machine learning in general, which can be grouped into two main categories: internal validation and external validation. Internal validation is by far the most used technique in text difficulty research for assessing the generalizability of trained models. There exist several variants of this technique, but the most common approach is the K-fold cross-validation. In this approach, a dataset is split into $K$ folds, and each fold is used in turn to validate the trained models (more details about cross-validation is discussed later in Chapter 4).

The second approach relies on the use of external sources, e.g., already available annotated corpus or formula, for validating the generalizability of trained models. While convenient, this approach requires that the newly developed text difficulty model target the same domain on which the original readability was developed.

A third possible validation approach is to continuously check and validate the decisions made by text difficulty systems with actual readers or listeners; in this sense that the actual beneficiaries of such systems are kept in the loop. One way to implement this validation technique is through calibrating estimations of text difficulty with real-time sensory data, e.g., eye-tracking, or thermal camera.

To recap, in this section I outlined six principles for developing more accurate text difficulty assessment tools, including those for spoken and videotexts. I also highlighted key issues that need to be taken seriously when developing computational models for text difficulty assessment and offered some possible ways to mitigate their impact. These six principles will serve as the base for developing our computational models for L2 videotext difficulty assessment.

## 2.3 Chapter Summary

In this chapter, I provided a succinct historical overview of text difficulty research and showed how the field has progressed over the years to become an interdisciplinary field

of research with practical uses and applications in various domains. I also outlined major computational approaches for measuring text difficulty and discussed how the field of text readability has benefited from advances in text comprehension research and the development of sophisticated computational techniques in computational linguistics and machine learning. Later, I reviewed, the small yet growing, body of literature on assessing L2 written and spoken text difficulty. Finally, in the light of a discussion about the key limitations of computational assessment of text difficulty, I provided six principles for developing more reliable and valid assessment of text difficulty, which will serve as the base for developing our computational models for L2 videotext difficulty assessment.

In the next chapter, I turn to videotext complexity and interrogate what makes a videotext difficult for language learners. Before that I place my pursuit of developing automated measures for videotext difficulty on a conceptual footing.

# Chapter 3

# Understanding Second Language Videotext Difficulty

## 3.1 Introduction

In the previous chapter, I reviewed the literature on text difficulty and outlined different computational methods for assessing difficulty in written and spoken texts. In this chapter, I turn to the main focus of this thesis, videotext difficulty, and interrogate what makes videotext difficult for language learners. In the first section, I review and discuss fundamental concepts and principles from the literature on (visual) narrative comprehension to build a conceptual understanding of the cognitive processes underpinning videotext comprehension. Based on this understanding, I then provide a working definition of videotext difficulty and describe its interaction with videotext complexity and comprehension, leaving aside issues related to task difficulty and contexts of learning. In the second section, I scrutinize L2 video studies published in the last two decades to ascertain how videotext difficulty was investigated. In the last section of the chapter, I allocate and synthesize empirical evidence of videotext complexity from diverse fields of study.

## 3.2 Theoretical Basis for Videotext Complexity: Laying the Foundations

For explaining a complex and multifaceted phenomenon as text comprehension, there is no a single theory but many (see Mcnamara & Magliano, 2009, for a comprehensive review). Consequently, the theoretical landscape is complex (Rayner & Reichle, 2010). While the majority of contemporary models of comprehension have primarily focused on printed texts, it is assumed, at least implicitly, that the assumptions made by these models in principle generalize to the comprehension of other types of texts, including videotext (Ferstl, 2018; Mcnamara & Magliano, 2009). This focus on printed texts can be partially attributed to the fact that a written text is easily controlled, manipulated, and analyzed than any other type of text (Mcnamara & Magliano, 2009).

In essence, theories of comprehension are universally concerned with the emergent mental representations that comprehenders generate when reading or listening to a text. The comprehender comes to specific mental representation from information in the text, information that is related to the text, and the inferences that he/she generates (Mcnamara & Magliano, 2009). Models of text comprehension assume that comprehension is the result of the successful understanding of words, sentences, and the relations between different parts of the texts. In a comprehensive review, Mcnamara and Magliano (2009) evaluated and compared seven prominent comprehension models: the *Construction-Integration model* (Kintsch, 1988, 1998; van Dijk & Kintsch, 1983), *Structure-Building Model* (Gernsbacher, 1990, 1997), *Resonance Model* (Albrecht & Myers, 1995; Albrecht & O'Brien, 1993), *Event-Indexing Model* (Zwaan & Radvansky, 1998), *Causal Network Model* (Young Suh & van den Broek, 1989), *Constructionist* (Graesser, Singer, & Trabasso, 1994), and *Landscape Model* (van den Broek, Risden, Fletcher, & Thurlssssow, 1996). The primary finding of this review is that "the current models of comprehension are not necessarily contradictory, but rather cover different spectrums of comprehension processes. Further, no one model adequately accounts for a wide variety of reading situations that have been observed, and the range of comprehension considered thus far in comprehension models is too limited" (Mcnamara & Magliano, 2009, p. 298).

Of the seven models reviewed in Mcnamara and Magliano (2009), three general models, *the Construction-Integration model*, *the Structure-Building model*, and *the Landscape model*, are succinctly discussed below. The three models are chosen because they encompass common (high-level) processes involved in text comprehension. It should be noted

that low-level processes germane to language understanding, including lexical decoding and syntactic parsing, are not accounted for in these models, or any other models of text and discourse comprehension in general. This is because the field of text comprehension "picks up after basic language understanding" (Mcnamara & Magliano, 2009). Our overarching purpose of reviewing these global or generic theoretical frameworks here is to provide an overview of the component subsystems of text comprehension and become acquainted with the assumptions made by these frameworks. Later, I review some recent models, which are built upon early models, whose objective is to provide a more exhaustive account of comprehension.

### 3.2.1 General Theories of Text Comprehension

One of the earliest and perhaps most influential models of text or discourse comprehension is the *Construction-Integration* (CI) model. The CI was first proposed by Kintsch in 1988 and then later expanded in his book (Kintsch, 1998). The CI has been further reformulated over the years in multiple books and an uncountable number of empirical studies (Kintsch, 1988, 1998; van Dijk & Kintsch, 1983), hence it is very challenging to describe all the assumptions postulated by the CI model, and it only suffices here to discuss its key premises.

A key concept in the CI model is the multiple levels of representation (Kintsch, 1988, 1998; van Dijk & Kintsch, 1983). In particular, the model assumes that three levels of mental representations are constructed in the reader's mind when reading a sentence: the surface structure, the propositional textbase, and the situation model. The surface level represents the words and phrases in the text and their syntactic relations. Text cohesion is represented at this level. The second level is the textbase representation and it is represented in terms of propositions. A preposition is a single unit of possessing and it consists of a predicate and one or more arguments. Kintsch and Van Dijk (1978) borrowed the prepositional notation from logic as a short-cut for representing ideas and their interrelations in a text, discourse, or scene regardless of their perceptual format, e.g., verbal, visual, or written. In representing ideas propositionally, neither grammatical structure nor word choices play a role, but only the semantic meaning (Ferstl, 2018).

The third level of representation is the situation model or what knowledge a reader or listener constructs from extracting new information from the text and relating it to prior knowledge and experiences. Since its introduction in van Dijk and Kintsch (1983), the

situation or mental model has attracted much attention and has been extensively investigated in the discourse comprehension literature (Zwaan & Radvansky, 1998). A situation model includes inferences made by the reader which go beyond the concepts explicitly mentioned in the text. Inferences that connect different elements of the story are called *bridging inferences* and virtually all discourse theories emphasize their role in constructing a coherent mental model (Mcnamara & Magliano, 2009).

Cognitive researchers have investigated the interplay between situation models, text cohesion, and inference generation. Cohesion and coherence are two key constructs in the discourse comprehension literature but the boundary between them is often blurred. A common distinction made between the two concepts (McNamara et al., 2014) is that cohesion refers to explicitly marked relationships such as repeated lexical words and pronouns (Halliday & Hasan, 1976; McNamara, Louwerse, McCarthy, & Graesser, 2010) while coherence generally refers to the psychological phenomenon of how a text comprehender understand relationships between different segments of the text (Givón, 2001). To put it differently, cohesion lies in the text whereas coherence lies in the reader's mind.

There is, however, no real consensus in the literature over this distinction, and the two concepts are argued to be interrelated. But researchers generally agree that "cohesion constitutes the explicit signals for managing coherence, and coherence is the broader notion of all relations, implicit or explicit, in the text or the mind" (Green, 2015, p. 327). Thereby, cohesion can be quantified and measured while coherence is hard to directly measure given that its dependence on the comprehender's world knowledge, level of education, and so on (McNamara et al., 2014).

Notwithstanding some criticisms, the CI model has inspired and influenced the development of several contemporary models of text comprehension, such as *the Landscape Model* (van den Broek et al., 1996), *the Event-Indexing Model* (Zwaan & Radvansky, 1998), and *the Reading Systems Framework* (Perfetti & Stafura, 2014).

The *Event-Indexing Model* was built upon the concept of situation model but it further elaborated the concept to provide a comprehensive account of how situation models are constructed via tracking various aspects of narrativity in texts. A focal assumption in this model is that comprehenders establish situation models along five dimensions, time, space, causality, motivation, and agent, and concurrently monitor and update current situation models as changes in the five dimension occurs. Accordingly, the *Event-Indexing model* is most applicable to dynamic texts or texts in which events are depicted concurrently in both space and time, as in film or videotext.

Another attempt to develop a general theory of comprehension regardless of text medium was made by Gernsbacher in his *Structure-Building model* (Gernsbacher, 1990, 1997). The model postulates that there are common cognitive abilities involved in the comprehension of various media, e.g., text, picture, or video. The model describes comprehension in terms of three key processes: (1) laying a foundation or a mental representation of an idea in a text, (2) mapping new information onto the foundation, and (3) shifting to new foundations or structures when there is a discrepancy between old and new information and the new information cannot map onto the existing foundation. According to this model, laying a foundation is an iterative process that operates whenever the comprehender encounters new information at the beginning of a story, scene, chapter, or even sentence. Compared to the mapping and shifting process, lying the foundation is assumed to be more resource-demanding (Mcnamara & Magliano, 2009). This assumption has been supported by empirical studies reporting slower processing times associated with reading the first sentence of a paragraph (e.g., Glanzer, Fischer, & Dorfman, 1984) or the initial picture of a picture story (Gernsbacher, 1983).

While the theories reviewed earlier, and almost all discourse comprehension for that matter, are developed for explaining high-level mechanisms involved in discourse comprehension, some recent attempts have been made to provide a more comprehensive unified model of comprehension, one that potentially can explain the full scope of comprehension processes (Mcnamara & Magliano, 2009). Of course, developing a theory that is being parsimonious while also capable of adequately explaining all the different processes involved and their interactions for various types of texts, readers, contents is not trivial.

Recently, there have been some attempts to synthesize and integrate all the various processes involved in narrative comprehension into a single account. One such recent attempt is the generic theory of comprehension put forward by Aryadoust (2019). The theory draws upon theoretical work from both reading and listening literature in order to provide a coherent and plausible explanation of auditory and written text comprehension. In particular, the theory "highlights the role of perception and word recognition (underresearched in reading research), situation models (missing in listening research), mental imagery (missing in both streams), and inferencing" (Aryadoust, 2019, p. 1). While the theory provides a plausible postulation for comprehending both written and auditory inputs, it nonetheless does little to explicate how visual information is extracted, processed, and integrated. Additionally, the theory presupposes a linear prorogation of information, where a listener processes and constructs meaning from aural input in an incremental fashion. This view of listening has been the subject of scrutiny, and it is now

believed that listening is a bidirectional process, involving both bottom-up and top-down processes (Buck, 2001; Rubin, 1994; Vandergrift, 2004; Vandergrift & Goh, 2012). This leads us to review next an emerging body of literature that seeks to describe how people comprehend sequential visual narratives (e.g., picture books, comics, and films), and how during this process perceptual and comprehension mechanisms are coordinated.

### 3.2.2 Visual Scene Perception and Sequential Narrative Comprehension

Owing partially to their inherent complexities, there is remarkably less theoretical and empirical research on audiovisual texts compared to written or spoken texts; but rather compartmentalized literature across different domains of cognitive sciences (e.g., memory, scene perception, event perception, and narrative comprehension). Recently, visual narrative scientists began to address this issue by proposing frameworks that integrate the various processes engaged in comprehending visual narratives (picture stories, comics, and film) into a coherent theory for both static (e.g., comics) and dynamic visual narratives (e.g., films) (e.g., Cohn, 2013, 2016b, 2020; Kendeou et al., 2019; Loschky et al., 2018, 2020; Schnotz, 2013). Comprehensive theoretical frameworks not only help explain how these seemingly different processes are coordinated to support the perception and understanding of complex visual narratives but also help to prevent in the field of visual narrative from fragmentation, as occurred, for example in reading research, where several models account for aspects of reading comprehension, such as word identification, syntactic parsing, discourse representations and so forth (Rayner & Reichle, 2010).

Work in this direction is stimulated by the postulation that human cognitive mechanisms underpinning comprehension are independent of the modality through which the information is received (Magliano, Radvansky, & Copeland, 2007). Here, I shall discuss key insights from three such models, namely the Visual Language Theory (VLT) (Cohn, 2013), the Parallel Interfacing Narrative-Semantics Model (PINS) (Cohn, 2020), and the Scene Perception and Event Comprehension Theory (SPECT) (Loschky et al., 2018, 2020).

Cohn (2013) developed Visual Narrative Grammar (VNG) to characterize how sequential narratives are processed and comprehended. In VNG, Cohn (2013) proposed that semantic cues within images in a sequence map to narrative roles organized within hierarchic constituents at a "discourse" level of meaning, analogous to the way that syntactic categories organize words into constituents at a sentence level. Sequential narratives are

conceived to share important features with language in the sense that they both have semantic content (conveyed by words and/or pictures) and sequencing rules (e.g., grammar) that govern how that content can be combined to form meaningful sequences. Moreover, both text-based and sequential narratives involve conveying a sequence of events that comprise a hierarchically structured plot (e.g., structured around goal plans; (Trabasso & Nickels, 1992). As such, VLT assumes that there are shared cognitive and brain systems that support the processing of both linguistic discourse and sequential narratives. As VLT was developed for wordless visual narrative wherein visual language is dominant (e.g., comics), it may not apply to other types of visual narratives (e.g., films) wherein verbal language is the primary carrier of meaning (but see Cohn, 2016a, 2016b).

Loschky et al. (2018) proposed the Scene Perception and Event Comprehension Theory (SPECT), an integrative theoretical framework for understanding both visual narrative perception and comprehension. In particular, it combines concepts from theories in the domains of scene perception (Henderson & Hollingworth, 1999; Oliva & Torralba, 2006; Wolfe, Võ, Evans, & Greene, 2011), event perception (Kurby & Zacks, 2008; Zacks, Speer, Swallow, Braver, & Reynolds, 2007), and narrative comprehension (Gernsbacher, 1990; Mcnamara & Magliano, 2009). SPECT describes how perceptual processing combines with mental model construction throughout visual narrative comprehension. It distinguishes front-end and back-end cognitive mechanisms. Front-end processes occur during single eye fixations and are comprised of attentional selection and information extraction, whereas back-end processes occur across multiple fixations and support the construction of event models, which reflect an understanding of what is happening now in a narrative (stored in working memory) and over the course of the entire narrative (stored in long-term episodic memory). More importantly, front-end and back-end processes interact with each other in a bi-directional fashion. Moreover, these two mechanisms are not equivalent to top-down and bottom-up processes, as characterized in numerous comprehension models. While the SPECT model was built upon decades of theoretical developments in general cognition and its subsystems and indeed provides a sensible and comprehensive account of how perceptual and narrative comprehension processes are integrated, Loschky and associates noted that "SPECT is not yet a formally complete cognitive model of visual narrative processing, primarily because many assumptions of the model remain to be empirically validated" (Loschky et al., 2020, p.29). This remark extends to other recently proposed models (e.g., Kuperberg, 2020) in this growing literature of visual narrative.

Integrating the concept of narrative structure from the VLT as well as drawing upon several other concepts from discourse comprehension literature, Cohn (2020) developed the Parallel Interfacing Narrative-Semantics Model (PINS Model), a processing model of visual narrative comprehension. The essence of the PINS model is parallel processing which involves two representational levels of semantics and narrative structure. According to this model, semantic processing alone is insufficient for understating visual narratives. Viewers need to construct a narrative structure that runs parallel to semantic processing. How this narrative structure is constructed is governed by the VNG (Cohn, 2013).

### 3.2.3 Bringing Order Out of Chaos: A Working Definition of Videotext Difficulty

Despite its central role in language teaching, learning, and assessment, the construct of 'difficulty' remains elusive and hard to pin down. Indeed, the term 'difficulty' is often used alongside or in place of terms such as complexity, comprehensibility, learnability, teachability, and accessibility. Interestingly, complexity has also been perceived positively and negatively in second language research. In learning a language feature, for example, complexity is often perceived as problematic, but some degree of complexity is beneficial (E. L. Bjork & Bjork, 2011; R. Bjork, 1994; Suzuki, Nakata, & Dekeyser, 2019). In language production research, complexity is a desirable and sought-after indicator of proficiency, but receptively complex language can reduce readability and intelligibility. It is therefore no surprise that the literature is depleted with numerous definitions and different operationalizations of complexity (see Bulté & Housen, 2012; Housen, De Clercq, Kuiken, & Vedder, 2019; Pallotti, 2015, for various approaches to linguistic complexity).

Pallotti (2015) remarked that complexity has figured in applied linguistics research to denote at least three meanings:

1. Structural complexity, a formal property of texts and linguistic systems having to do with the number of their elements and their relational patterns;

2. Cognitive complexity, having to do with the processing costs associated with linguistic structures;

3. Developmental complexity, i.e. the order in which linguistic structures emerge and are mastered in second (and, possibly, first) language acquisition (p.118).

Pallotti (2015) asserted that the three meanings refer to distinct and analytically separable constructs, "even if they were found to correlate very strongly, this would not mean they are three facets or names of one single construct" (p. 119). Advocating for a simple view of linguistic complexity, Pallotti (2015) suggested restricting the construct to structural complexity only, and explicitly excluded "issues of cognitive cost (difficulty) or developmental dynamics (acquisition)" (p. 117). R. DeKeyser (2016) remarked that narrowly defining difficulty in terms of the inherent structural complexity of the language would eliminate useful discussions about other sources of difficulty, e.g., contextual and subjective difficulty.

Researchers have also attempted to detangle all the different sources of linguistic complexity (Bulté & Housen, 2012; Pallotti, 2015) and difficulty (R. DeKeyser, 2016; Housen & Simoens, 2016) and delineate their interactions in comprehensive frameworks or taxonomies. By choice, these taxonomies vary in their granularity and specificity of what linguistic complexity or difficulty entails in the realm of the second language. While elaborate taxonomies of such nature may indicate thoroughness and a clear sense of how the various language systems or subsystems interact, it is easy to imagine such taxonomies fall apart when tested empirically in a real, complex world.

In this thesis, I make clear distinctions between three seemingly related, but conceptually distinct and operationally separable constructs: videotext difficulty, complexity, and comprehension. Specifically, videotext difficulty is understood as *an increase in the attentional and cognitive resources demanded for successful understanding of videotext.* Therefore, videotext difficulty always "reflects rather than creates complexity" (Rescher, 1998, p. 17).

On the other hand, videotext complexity subsumes two broad notions of complexity: absolute complexity and relative complexity (Miestamo et al., 2008). Absolute complexity refers to the inherent characteristics of videotexts that make processing and comprehending a videotext difficult regardless of language learners or tasks. In this sense, absolute complexity may also be challenging for L1 viewers. On the other hand, relative complexity is the relative ease or difficulty of a videotext for an individual learner, for example, as a result of the attentional, memory, and other information processing demands imposed by the videotext, hence, a videotext can be difficult for a language learner may not be so for another. In this sense, the videotext difficulty is the resultant of the interaction between absolute and relative complexity. A sketch of this relationship is depicted in Figure 3.1.

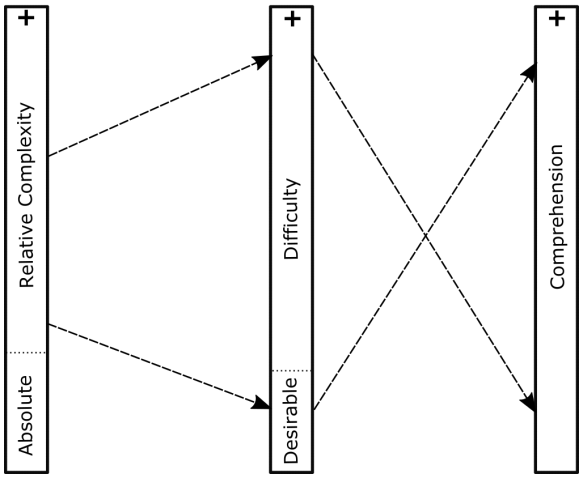



FIGURE 3.1: The Relationship between Videotext Complexity, Difficulty, and Comprehension

While an increase of videotext complexity often leads to the perception of difficulty, the relationship between difficulty and comprehension is not symmetrical. That is, videotext difficulty does not always lead to imperfect comprehension. Indeed, some degree of difficulty is desirable and may help learners to better understand the text, for example, through inducing the viewers to pay more attention and exert additional effort to the understanding of the videotext (E. L. Bjork & Bjork, 2011; R. Bjork, 1994; Suzuki et al., 2019). But, if a videotext is far beyond language learner's ability, it is conceivable that their comprehension of the text would be superficial—identifying the main gist of the videotext, mostly from visual imagery. Most importantly, difficulty should not be treated as a binary variable—texts are either easy or difficult—but rather a continuous variable.

## 3.3 Video in Second Language Research

As an initial step towards understanding L2 videotext difficulty, I reviewed [1] video studies published between 2000 and 2020 in the broader field of Applied Linguistics. Specifically, I wanted to answer the following two questions:

- Q1: how videotexts have been researched in Applied Linguistics?

[1] I began searching for and collecting studies for this review in early June 2019 and kept doing so until December 28th, 2020.

- Q2: has videotext difficulty been a primary variable in past research? if not, how L2 researchers have determined or controlled the difficulty of videotexts in their research?

Although I did my search of relevant studies in the Web of Science databases which, of course, returned hundreds of entries, I restricted my scope of this preliminary review to studies published in some reputable Applied Linguistics journals, e.g., *Foreign Language Annals, Modern Language Journal, System, International Journal of Listening, Language Testing, Language Assessment Quarterly, TESOL Quarterly, English for Specific Purposes, Language Learning & Technology, Language and Communication Quarterly, International Journal of Language Studies, ReCALL, Computer Assisted Language Learning, and The Language Learning Journal.* To begin, I searched for relevant studies using the keyword "listening" in conjunction with one of the following terms: "video", "videotext", "TV", "audiovisual", "audio-visual", "multimedia", and "lecture". After initial screening, I realized that L2 videos have been used in areas other than listening (e.g., learning L2 'vocabulary', 'grammar', and 'reading'), so I expanded my search to include those studies.

After excluding irrelevant studies (e.g., reviews, reports, meta-analysis), there were in total 77 published empirical studies related to the use of video in L2 research, as shown in Figure 3.2. These studies have essentially centered, to a varying degree, around five research areas[2]; *advance organizers, subtitles and captions, visual processing, strategy use and instruction, assessment*, and *learner/teacher perception.* I briefly describe the key findings within each research area.

### 3.3.1 Areas of Research

#### 3.3.1.1 Caption and Subtitle

Video caption and subtitles (n=21) is an area of research in which the bulk of past work has been carried out. The key question in this line of research is whether video captions and/or subtitles facilitate video understanding, vocabulary acquisition, and language development (e.g., Winke, Isbell, & Ahn, 2019). Generally speaking, findings of previous studies have endorsed the use of L2 captions in areas such as aural word recognition, and video content comprehension and recall (e.g., Guichon & McLornan, 2008; Montero Perez,

[2]Very few studies were concerned with language perception towards watching videos (one study) and the development of video-based tools (two studies by the same authors). Therefore, they were not considered as a research area.

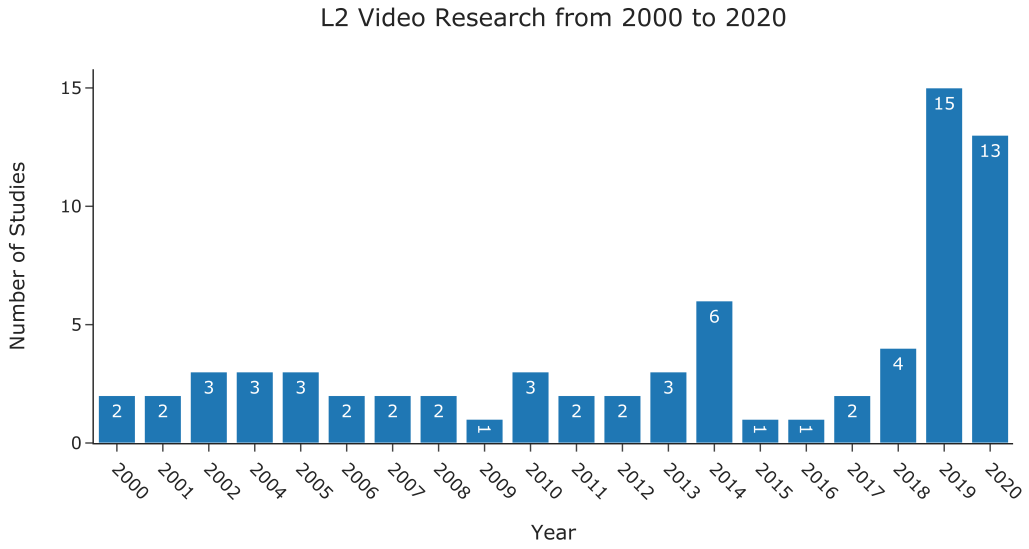



FIGURE 3.2: L2 Video Research from 2000 to 2020

Peters, & Desmet, 2014; Winke, Gass, & Sydorenko, 2010; J. C. Yang & Chang, 2014). Because speech is fleeting, captions are conceived to aid language listeners in distinguishing linguistic units (e.g., words, morphemes, grammatical components) and mapping aural speech to individual words and phrases (Winke et al., 2019). However, it has also been suggested that language learners may read the captions at the expense of listening to the speakers, thus related gains in video content comprehension and recall are more likely due to reading comprehension, not to viewing comprehension precisely (Vandergrift & Cross, 2014). Indeed, Yeldham (2018) investigated this claim by reviewing the results of some past studies on video captions. The findings of his investigation revealed that less proficient language learners tend to mainly read captions rather than listening to videotext, while more-proficient learners tend less to read captions and utilize more cues in the videotexts (speakers and visual).

#### 3.3.1.2 Video-Based Assessment

In the language assessment literature, the use of videotext in creating language tests has received some notable attention (e.g., Batty, 2014; Gruba, 1993; Pardo-Ballester, 2016; Suvorov, 2014a; Wagner, 2010a, 2014). Proponents of their use in testing often cited that videotexts are more ecologically valid than audio-based tests due to the multimodal nature of videotexts which resembles everyday listening. Early research in this area was mainly focused on examining whether listening comprehension is affected by how information is presented to the listeners, i.e., via video or audio. The findings of these studies

were, by and large, inconclusive and contradictory. Some studies found videotext to be conducive to text comprehension (Wagner, 2010b, 2013); other studies found no significant effect (Batty, 2014; Coniam, 2001; Gruba, 1993; Londe, 2009) or even had a debilitating effect (Suvorov, 2009). These inclusive findings are not surprising, however. Comparative studies of this nature are invidious because they ignore the remarkable differences between the two different types of texts, audios vs videos, let alone differences within individual texts in each medium. However, in recent years, researchers seem to abandon comparative studies and focus more on single-modality assessment studies (e.g., Batty, 2020).

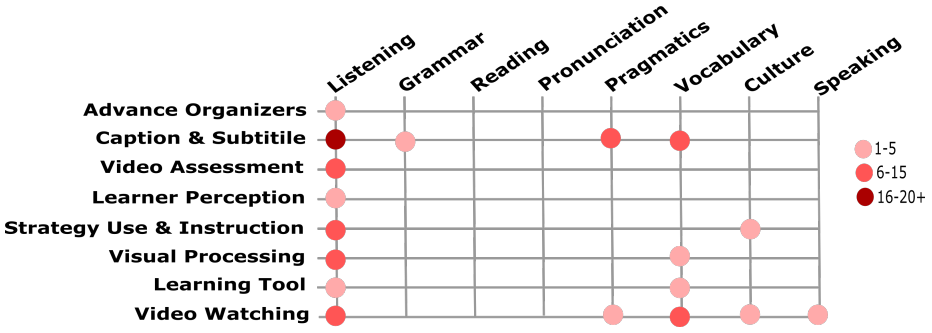



FIGURE 3.3: Areas of L2 Video Research from 2000 to 2020

#### 3.3.1.3 Visual Processing

Turning to visual processing research, several studies have examined how L2 listeners deal with the visual content of videos and whether contextual visual cues help or hinder video comprehension. Ostensibly, visual cues in videos provide contextual support for language learners but the findings of previous research were inconclusive. For example, Gruba (2004) and Cross (2011) similarly observed that when language learners watch television news they switch their attention between the audio and visual channels largely due to an excessive cognitive load induced by processing information coming from both channels. In Sueyoshi and Hardison (2005), intermediate and advanced learners of English improved their comprehension from watching a lecture in which the speaker's gestures and facial expressions are shown. A key challenge in this line of research is understanding how and under what conditions visual imagery helps or impedes L2 listening comprehension. Visual imagery is generally understood as to be of two types; content and context visual (Bejar, Douglas, Jamieson, Nissan, & Turner, 2000; Ginther, 2002). Context visuals provide information about the context of the verbal exchanges, such as who are the participants and where the conversation is happening. The main role of context visual is to set the scene for the verbal exchange. Content visual, on the other hand, is an

essential component of the text and they may reiterate, illustrate, organize, and supplement information in the verbal stimulus (Bejar et al., 2000). Of course, such depiction of visual imagery is very simplistic and does not consider the complex interactions between verbal and visual language. In the absence of a sufficient understanding of what impact different forms of visual imagery have on second language listening, Gruba recommended that visual imagery "is better thought of as integral resources to comprehension whose influence shifts from primary to secondary importance as a listener develops a mature understanding of the videotext." (Gruba, 2004, p.51)

#### 3.3.1.4 Strategy Use and Instruction

Videotext has also been the focus of some instruction research where the primary focus was on examining the effects of different strategy instruction on videotext comprehension. Compared to other L2 video research, there are relatively very few studies in this line of research. For instance, Seo (2005) and Cross (2009) explored the impact of strategy instructions on improving learners' comprehension of news videotexts and both studies could not confirm the positive impact of strategy instruction as learners in the control group showed signs of improvement as well.

#### 3.3.1.5 Advanced Organizers

As for advanced organizers, i.e., materials that help listeners activate context knowledge and prior knowledge (Ausubel, 1960), a few studies have shown that advanced organizers appear to be effective in aiding L2 learners to comprehend video content. Regardless of what modality it is presented in, written (Jafari & Hashim, 2012), aural (Chung & Huang, 1998) or an aural advance organizer accompanied by pictures (Wilberschied & Berman, 2004), the use of advance organizers seems to be beneficial for L2 videotext comprehension: but how helpful it is with more difficult videotext is still unknown.

### 3.3.2 Videotext Difficulty in L2 Video Research

My examination of past research showed that 1) videotext difficulty has never been studied explicitly as a primary research variable and 2) it was not considered as a confounding variable—whose impact might influence or moderate variables under investigation—in most studies (70%). Indeed, there was a tendency until recently to downplay the role

of videotext difficulty. I found this slightly odd because videotext difficulty impact language learners' watching behaviors, emotions, and ability to comprehend the text. In my view, not controlling for videotext difficulty in second language research, regardless of the objective of the study and language proficiency of the participants, leads to conflicting results, or even worse, false conclusion.

Research that has accounted for the variability in videotext due to difficulty did so in several ways (see Figure 3.4). Common approaches for judging videotext difficulty were teachers' subjective evaluation (e.g., Seo, 2005), the use of L1 readability formulas (e.g., Mestres & Pellicer-Sánchez, 2019), lexical coverage (e.g., Y. Feng & Webb, 2019; Puimège & Peters, 2019), pilot testing (e.g., Gruba, 2006), or simply adopting already published materials (e.g., Li, 2015). These approaches suffer from some drawbacks. For example, the use of readability formulas and reading measures for gauging the level of difficulty in videotexts ignores the unique nature of videotexts and does not account for the processes involved in understanding audiovisual materials. The reliance on off-the-shelf language learning textbooks or TV shows is problematic because criteria used by publishers in choosing or designing videotexts are normally not available to the public. Therefore, L2 researchers cannot make an informed decision. More recently, researchers relied on lexical frequency lists as a sole indicator or in combination with speech rate to determine the difficulty in videotexts, but it is not clear how researchers decided to use, for example, words from K. Perhaps a less problematic approach involves pilot-testing videotexts with a small group of learners or asking a committee of experts to check the appropriateness of videotexts before running the main study (e.g., Gruba, 2006; Seo, 2005) But again, subjective evolution of such nature may render results that are not generalizable to other contexts.

To recap, videotext has attracted considerable attention, especially of late. While researchers have investigated different uses and aspects of videotexts, it was video caption and subtitle research that breathed a new life into L2 video research. Indeed, out of 15 studies published in 2019, 10 studies were concerning the use of video captions or subtitles. It was also found that information about the videotext and its content was either not provided or provided but vaguely described. Nor were there any detailed descriptions of language and visual imagery used in the videotext. Unfortunately, in the absence of such information, reliable inferences and conclusions cannot be readily made from these studies.

As far as videotext complexity or difficulty is concerned, there has been no systematic investigation directed towards understanding videotext difficulty and complexity in L2.

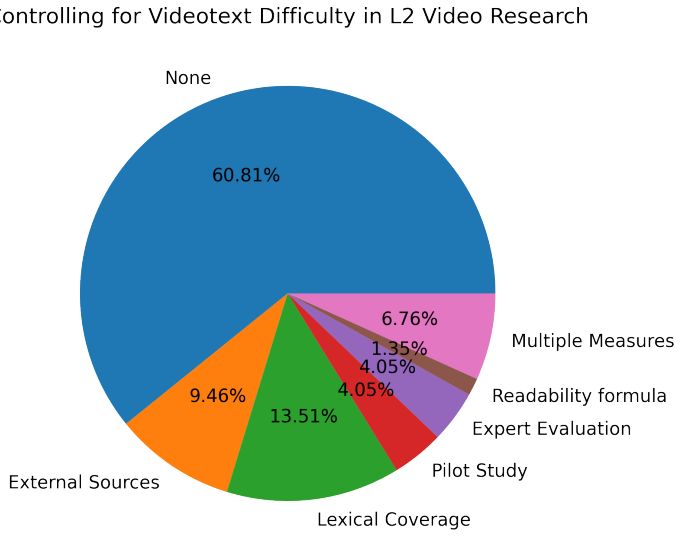



FIGURE 3.4: How Videotext Difficulty was Assessed in L2 Video Research

What is more, videotext difficulty is rarely conceived as a confounding factor that might influence the outcomes of the studies and hence their generalizability and reproducibility. By way of illustration, research on the effects of video captions has generally adopted an experimental research design where one group of language learners watch a captioned video and the other group watch a video with no caption. If presented with linguistically difficult videotexts, learners are most likely going to rely on video captions to understand the videotext. And by the same token, if the videotext is easy, and on a topic that the learner is interested in, it is more plausible that the language learner pays less attention to caption and more attention to the videotext (Yeldham, 2018). In a recent caption study in which an eye-tracking technology is used, Gass, Winke, Isbell, and Ahn (2019) concluded that "there may be a zone within which captions are most useful, perhaps centered where the content difficulty level is not too far above or below a learner's ability level" (p. 88). The same leniency for considering the impact of videotext difficulty is also seen in many studies in the other research areas, including language assessment and vocabulary learning.

As might become clear by now that my review of the literature was not very helpful in ascertaining what makes a videotext difficult for language learners. There is, as such, a clear void in the literature that needs to be filled. To build an understanding of videotext difficulty, I review and discuss key principles and concepts from the field of discourse comprehension and visual narrative comprehension.

## 3.4 Empirical Evidence of Complexity

Definitional and operational issues aside, it should be noted that videotext difficulty is not a unitary construct but rather a multifaceted and multilayered one that is the resultant of the interaction of many factors pertaining to the text, task, and learner. In this section, I only review and synthesize empirical evidence of verbal and visual complexity from various research fields and domains. In examining empirical evidence, what theory they lend support to and from which research discipline they originate bear less importance. What matters is there is abundant evidence supporting a particular claim.

### 3.4.1 Verbal Complexity

Undeniably, language is the primary bearer of information in videotext. For language learners, who are still at the early stages of their language learning development, it is very reasonable to assume that many of the difficulties they encounter arise from their limited knowledge of the target language. A large body of research has interrogated several aspects of language that seem to be more problematic for language learners. In this thesis, I use the term "video verbal complexity" as an umbrella term to refer to all of these aspects, including issues related to acoustic, phonological, lexical, semantic, syntactic, and discoursal complexity. Although there is a burgeoning literature covering each of these aspects, I only explore key indicators of complexity that are directly relevant to the investigation pursued in this thesis. For a broader review of factors, including affective and contextual factors, affecting listening comprehension in L2, I refer the reader to Vandergrift and Goh (2012).

#### 3.4.1.1 Acoustic and Prosodic Complexity

Language learners need to have a working knowledge of the acoustic and prosodic system of the target language to mitigate difficulties due to variability in speech, e.g., acoustic dissimilarity. This section reviews research particularly germane to acoustic and prosodic characteristics of spoken language that impair speech processing in L2, including, for example, accented speech, speech rate, articulation rate, pauses, and speech disfluencies. The review draws on L1 studies where there is a gap in the available L2 literature. Summary of the key findings is presented in Table 3.1

#### 3.4.1.1.1 Accented Speech

Even in one's first language, listening to an unfamiliar accent can be challenging and listeners may experience difficulty and a delay in recognizing unfamiliar spoken utterances (Major, Fitzmaurice, Bunta, & Balasubramanian, 2005; Weil, 2001). To cope with this situation, listeners must adapt quickly to the variations in the linguistic form. The difficulty in the adaption increases when the listener's and speaker's language variety diverge phonologically and phonetically. Therefore, it is logical to say that the effect of accent is further exaggerated in the case of second language listeners, especially for those at the beginning level of language proficiency. It is also understood that L2 learners are more likely to be less familiar with different accents in their L2, especially uncommon or regional ones. Previous studies investigating the effect of unfamiliar accents on comprehensibility have recurrently showed empirical evidence indicating a direct relationship between unfamiliar accent and speech comprehensibility (Adank, Evans, Stuart-Smith, & Scott, 2009; Floccia, Butler, Goslin, & Ellis, 2009; Major et al., 2005; Smith & Bisazza, 1982).

In one of the earliest studies, Smith and Bisazza (1978) or (Smith & Bisazza, 1982), for instance, presented L2 learners of English from Hong Kong, India, the Philippines, Japan, Taiwan, and Thailand with identical passage spoken by non-native (Indian and Japanese), and native (American) speakers of English. The study found that the United States speaker was significantly easier to comprehend than the Japanese speaker, who in turn was significantly easier to comprehend than the Indian speaker. Similarly, Major et al. (2005) explored whether L2 learners experience difficulty comprehending regional (Southern American English), ethnic (African American English), and international dialects of English (Australian English, and subcontinental Indian English) more than Standard American English. After listening to lectures delivered in one of these dialects, L2 learners took a listening comprehension test. The results showed that L2 learners scored significantly lower on the tests hearing ethnic and international dialects of English, but no significant difference was found on their scores when hearing regional dialect and Standard American English. The finding of Major and colleagues' study suggest that L2 learner's familiarity with the speaker's accent seems to have an impact on their listening comprehension.

Recently, Ockey and French (2016) conducted a study to determine the impact of the strength of accent on listening comprehension of lectures. In their study, more than 21,000 TOEFL iBT test takers listened to monologic lectures narrated by English speakers (American, Australian, and British) who were selected based on a judgment of their strength of accent. The findings demonstrated that the strength of accent and familiarity do affect

listening comprehension, and these factors affect comprehension even with quite light accents. The same finding was also observed with L2 learners listening to interactive lectures (Ockey, Papageorgiou, & French, 2016).

By and large, the vast majority of research on the effect of speaker accent seems to suggest a direct relationship between accent familiarity, and speech comprehensibility and intelligibility. However, some studies failed to observe any significant effect of accented speech on L2 listening comprehension. For example, in her study, Abeywickrama (2013) tested the comprehension of Brazilian, Korean, and Sri Lankan English after listening to passages spoken in four different varieties of English. The findings of her study showed no impact of accented speech on learners' performance on the listening tests. One possible explanation for the contradictory findings of previous studies, as noted by Ockey et al. (2016), can be ascribed to the different definitions and operationalization of the construct adopted by researchers. Moreover, there seems to be a myriad of other factors that seem to work against learner's ability to comprehend an accented speech, including, for example, environment noise (Adank et al., 2009), speech rate (Matsuura, Chiba, Mahoney, & Rilling, 2014) and whether or not listening to an English speaker with whom they share language background (Kang, Thomson, & Moran, 2018; Major, Fitzmaurice, Bunta, & Balasubramanian, 2008).

#### 3.4.1.1.2 Speed of Delivery

It is widely acknowledged that fast speech makes processing, decoding, and recognizing spoken words much harder, and "when speech rate reaches a critical level, comprehension becomes all but impossible" (Renandya & Farrell, 2011, p. 53). For language learners who are still developing their bottom-up listening abilities, listening to a speech at a normal speed or even at a slow speed is generally perceived to be too fast by beginning language learners (Flowerdew & Miller, 1992; Graham, 2006; Renandya & Farrell, 2011). However, evidence from research on speech rate is so far inconclusive. Several studies have demonstrated that listening comprehension tends to decline when speech rate is fast (Brindley & Slatyer, 2002; Griffiths, 1990, 1992; Y. Zhao, 1997). For example, Griffiths (1990) explored the impact of three speech rates–200, 150, and 100 words per minute (WPM)–on listening comprehension. The results demonstrated a significant reduction in comprehension when passages delivered at speech rates of 200 WPM/ 3.8 Syllables per second (SPS), but comprehension scores on passages delivered at slow rates did not significantly differ from those delivered at an average rate. In a later study, Griffiths (1992) had 24 intermediate EFLs to listen to three stories delivered with a greater spread between

Table 3.1: Acoustic and Prosodic Complexity, their Impact & Supporting Empirical Evidence

| | **Hypothesized Impact** | **Empirical Evidence** |
|---|---|---|
| Speaker accent | Unfamiliarity with speaker accent may negatively impact spoken text comprehension | (Ockey & French, 2016; Ockey et al., 2016; Roy, Susan, Ferenc, & Chandrika, 2005) |
| Speech rate | Faster speech is more difficult than slower speech. | (Brindley & Slatyer, 2002) |
| Disfluencies | Filled and silent pauses may improve comprehension because they add extra processing cues for the listener. | (J. E. Arnold, Kam, & Tanenhaus, 2007; Cevasco & van den Broek, 2016) |
| Connected speech | Connected speech may affect the listener's ability to recognize and process spoken utterances. | (Henrichsen, 1984; Tuinman, Mitterer, & Cutler, 2012; S. Wong et al., 2017; S. W. Wong et al., 2017) |
| Rhythmic variability | Rhythmic variability may help listeners to predict when semantically important information will occur in speech. | (Allen & Hawkins, 1980; Cutler & Otake, 1994) |
| Intonation (pitch) | Changes in pitch may influence lexical processing of spoken input, especially in tone-based languages. | (Beach, 1991; Cutler, Dahan, & Van Donselaar, 1997; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991) |
| Phonotactic probability | Words with high phonotactic probability are more likely to be processed, segmented, acquired, and produced. | (Bradlow & Pisoni, 1999; Mintz, Walker, Welday, & Kidd, 2018; Vitevitch & Luce, 2004) |
| Phonological Neighborhood effect | Spoken word perception and production are affected by the degree to which the sound pattern of a given word is similar to other phonological patterns in memory | (Luce & Pisoni, 1998; Vitevitch, 2007) |

speech rates; slow (approximately 127 WPM), average (approximately 188 WPM), and fast (approximately 250 WPM). The results showed that a more significant difference between comprehension at a slow and fast rate. Instead of manipulating speech rate, Y. Zhao (1997) allowed each participant to adjust the speed of their listening passage and observed that participants comprehended better when they had control of their speech rate, and improved comprehension was achieved by slowing down the speech rate.

However, other studies have found moderate or no significant effect of speed of delivery on listening comprehension. For example, Blau (1990) found recording a speech at normal speed (approximately 170 wpm) and slowing it to 145 wpm did not significantly improve the listening comprehension level of more advanced English majored Polish university students and less advanced Puerto Rican learners enrolling ESL courses. In a second experiment, Blau (1990) compared the effect of speed and adding pauses. The results, overall, suggested that speech with pauses was consistently the most comprehensible whereas the reduced speech rate was the least comprehensible, especially for more advanced language learners.

It is important to note that it is still unclear yet at what speech rate L2 listening becomes easier or harder, and researchers have proposed different threshold levels around which listening becomes harder (Chang, 2018). Y. Zhao (1997) noted that what speech rate was considered to be 'normal' in previous studies was chosen arbitrarily by researchers, which undermined the generalizability of their findings. Furthermore, Y. Zhao (1997) asserted that the perception of whether a speech is fast or slow varies from one learner to another, and differences in learning style and language proficiency also impact how speech is perceived.

Another methodological issue is that researchers have employed various measures for determining speed delivery, including word per minute (WPM) and syllables per second (SPS). Despite being commonly used, WPM has the shortcoming of not controlling for word length; that is, it is plausible that two utterances have similar WPM values, yet one could contain largely monosyllabic words while the other could contain multisyllabic words (Bloomfield et al., 2010; Griffiths, 1990, 1992). Although SPS takes account of variation in the number of syllables per word, it may indicate an articulatory rate as opposed to speech rate (Bloomfield et al., 2010). Articulatory rate, unlike speed rate, excludes silent intervals that exceed a certain threshold. This threshold aims to preserve silent intervals that correspond to articulatory gestures while eliminate those corresponding to pauses related to speech planning or disfluencies. The values of such thresholds, however, vary in literature ranging anywhere from 150–250 ms or longer (e.g. Tsao, Weismer, & Iqbal,

2006). While both speech rate and articulatory rate are concerned with speech production, speech rate provides a global measure of speed that incorporates pausing time. To fully capture speed delivery, Bloomfield et al. (2010) suggested combining a speech rate measure with an articulation rate calculated by excluding silent pauses, because pausing may facilitate comprehension.

Following Bloomfield and colleagues' recommendation and to sidestep the limitation in previous studies, Brunfaut and Révész (2015), and Révész and Brunfaut (2013) included three measures of speed of delivery in their studies: articulation rate (number of syllables per second excluding pause time), speech rate (number of syllables per second including pauses), and number of pauses (silent periods exceeding .25 s.) per second. However, neither of the two studies found the speed of delivery measures to contribute to listening difficulty. It is worth noting that the speech samples in both studies were relatively small (18 and 30 short passages).

The conflicting results of existing work may suggest that there are perhaps other confounding factors that make language learners perceive speech to be fast, including unfamiliar accent, audio quality, text difficulty, topic unfamiliarity, and whether a text is scripted or unscripted (Bloomfield et al., 2010; Chang, 2018). Matsuura et al. (2014) investigated the interaction between accented speech and speech rate and observed that reducing speech rate helps Japanese university students studying EFL in Japan to comprehend a less familiar English (Indian English).

#### 3.4.1.1.3 Disfluencies and Pauses

Speech, especially spontaneous speech, is rarely continuously fluent and is typically replete with disfluencies, such as pauses, hesitation, and false starts. These disfluencies occur up to six times per hundred words of speech (Tree, 1995). Historically, disfluencies in the speech were once thought to serve no function, but recent psycholinguistic research suggests that disfluencies serve rhetorical and communicative purposes, such as orienting the listener's attention (Collard, Corley, MacGregor, & Donaldson, 2008), providing clues about what the speaker is about to describe (Corley & Hartsuiker, 2003), and leading listeners to expect new information (J. E. Arnold, Fagnano, & Tanenhaus, 2003). Though few in numbers, studies on the role of silent and filled (e.g. um, uh) pauses on second language listening demonstrated that they aid listening comprehension, mainly because they allow for extra time for processing input (Blau, 1990; Buck, 2001). Pauses seem to benefit low-proficiency learners more than high-proficiency learners, who might find frequent pauses distracting (Watanabe, Hirose, Den, & Minematsu, 2008).

There is little research on the impact of disfluencies on L2 text comprehension. Previous research generally found that disfluencies seem to facilitate spoken text comprehension, especially for more proficient listeners (Bloomfield et al., 2010; Cevasco & van den Broek, 2016). Researchers such as Blau (1990) and Watanabe et al. (2008) provided empirical evidence for the positive role of filled pauses on improving L2 spoken text comprehension because they allow for more processing time (Buck, 2001). The role of disfluencies in L2 text comprehension. Researchers also found that the effects of filled pauses and silent pauses may depend on the listener's language proficiency, and beginner learners seem to benefit the most from pauses in speech than advanced learners.

#### 3.4.1.1.4 Rhythm and Pitch

Rhythm and pitch are two acoustical properties of sound that have been shown to impact speech processing, word recognition, and resolve syntactic ambiguity (Beach, 1991; Roncaglia-Denissen, Schmidt-Kassow, & Kotz, 2013). Rhythm is responsible for arranging speech into a pattern of recurrent temporal units or events and languages vary in their rhythmic organization. In particular, rhythm plays a critical role in speech comprehension as it helps listeners to "track, segment, anticipate, and focus attention on high yield aspects of the speech signal and to facilitate recognition and prediction of syntactic and semantic relationships among words and phrases in continuous speech" (Liss, Utianski, & Lansford, 2013, p. 5). The perception of rhythmic cues plays an important role in recognizing spoken language, especially in adverse listening conditions and it helps to resolve syntactic ambiguity (Roncaglia-Denissen et al., 2013). Pitch refers to the degree of highness and lowness of speech and it is the acoustic equivalent of intonation. In stressed-based language, like English, variability in pitch plays a critical role and it helps listeners to recognize words from speech. Infants, as young as 7 months can recognize a word if it is presented in a different amplitude but not in a different pitch (Singh, White, & Morgan, 2008).

In the context of second language listening, relatively little research has been directed towards understanding how rhythm and pitch may help L2 listeners. Révész and Brunfaut (2013) explored the role of several phonological factors, including phonotactic probability, frequency of elisions, rhythm, and pitch, on L2 listening comprehension difficulty but observed no significant role of these factors on L2 learners' listening comprehension. The researchers partly attributed this finding to the fact that all of the listening passages were narrated by the same speaker who spoke standard British English and read from

the scripted texts. To date, little is known about the role of suprasegmentals on videotext difficulty.

#### 3.4.1.1.5 Phonological Complexity

Spoken language is often characterized by assimilation as well as unclear articulation, and lexical units are not necessarily as clearly marked as in written texts. This means listeners are often forced to make inferences about word boundaries based on incomplete acoustic information (Stæhr, 2009). This lack of clarity of spoken language makes the ability to segment and recognize words in continuous speech essential for successful L2 listening comprehension. In this section, I discuss several aspects of language that have been shown to make spoken text perception more challenging, including connected speech, phonotactic probability, and phonological neighborhood density.

#### 3.4.1.1.6 Connected Speech

Connected speech or sandhi variation (e.g., can't and I'll) is a defining characteristic of spoken language and it is more prevalent in an informal speech than formal speech (Brown, 2012; Field, 2003). Connected speech has been claimed to negatively affect spoken text comprehension because the phonemic information missing in the aural input makes the recognition of forms and syntactic patterns more difficult (Rubin, 1994). In one of the earliest L2 studies on connected speech, Henrichsen (1984) compared the impact of sandhi variations on spoken text comprehension and found that texts that include sandhi variation are more challenging for second language listeners when compared to native speakers. This finding is expected because in processing speech listeners have a shorter time to process incoming input and hence it becomes more difficult for less proficient listeners to decode all incoming information.

Révész and Brunfaut (2013) found the effect of reduced forms on listening passage not a significant predictor of listening difficulty. However, the listening texts in Révész and Brunfaut (2013) were scripted and delivered by actors, which might have decreased item variation. In a recent study, S. W. L. Wong et al. (2015) explored the interactions among various phonological skills (spoken word discrimination, part-word recognition, phonemic awareness, vocabulary knowledge, and phonological memory) thought to underlie the perception of English reduced forms in Chinese undergraduate ESL learners. The results of regression analyses revealed that reduced form dictation significantly predicted generally spoken text comprehension. Other factors may also impact the perception of reduced forms in continuous speech such as the adversity of the surrounding environments

(S. W. L. Wong et al., 2015) and learner's native language which may place phonotactic constraints on what they hear (Ernestus, Kouwenhoven, & van Mulken, 2017).

#### 3.4.1.1.7 Phonotactic Probability

Phonotactic probability refers to "the frequency with which phonological segments and sequences of phonological segments occur in words in a given language" (Vitevitch & Luce, 2004, p.481). In English, for instance, initial /spr-/ is a possible phonotactic sequence, whereas /spm-/ is not (Crystal, 1994). Several empirical studies have confirmed the influence of phonotactic probability on L1 language processing and comprehension by infants (Mattys & Jusczyk, 2001; Mintz et al., 2018) and adults (McQueen & Pitt, 1998). More specifically, words with high phonotactic probability are more likely to be segmented, acquired, processed, and produced.

One method for computing the phonotactic probability relies on averaging unigram or bigram positional probabilities across a word, that is how often a certain segment occurs in a particular position in a word (Vitevitch & Luce, 2004). While there is a considerable body of research suggesting the influence of phonotactic probability in L1, there is only a handful of studies that investigated the role of phonotactic probability on L2 listening comprehension (Bradlow & Pisoni, 1999; Révész & Brunfaut, 2013). For example, in their study Révész and Brunfaut (2013), gauged the phonotactic probability for 18 transcribed passages in terms of mean positional segment frequency and mean biphone frequency and found no significant impact of phonotactic probability on L2 listeners comprehension. Révész and Brunfaut (2013) attributed this finding to the fact that all listening passages in their study were narrated by one speaker, who spoke standard British English and read from scripted texts, which resulted in low phonotactic variation. Consequently, the authors called for further research as there are very few studies that have explored the impact of phonotactic probability on L2 listening.

#### 3.4.1.1.8 Phonological Neighborhood Density

The phonological neighborhood refers to words that are different in one and only one phoneme (Vitevitch & Luce, 2016). Phonological neighborhood density is a measure of how many neighbors a certain word has (Yao, 2011). A considerable body of research has demonstrated that both spoken word perception and production are affected by the degree to which the sound pattern of a given word is similar to other phonological patterns in memory (see Vitevitch & Luce, 2016, for an extensive review). Specifically, words that are similar to many other words are processed more slowly and less accurately than words

that are similar to only fewer words (Luce & Pisoni, 1998; Vitevitch, 2007). While previous research has mainly concentrated on native language processing and production, there seems to exist no prior research investigating how phonological neighborhood impacts processing second language videotexts and to what degree the presence of phonological neighbors in videotext contribute to its complexity.

#### 3.4.1.2 Lexical Complexity

Over the past two decades, a growing body of language studies has interrogated the relationship between lexical complexity and the comprehension of written and spoken text (Brunfaut & Révész, 2015; Révész & Brunfaut, 2013), writing quality (Crossley & McNamara, 2012), vocabulary learning and development (Hashimoto & Egbert, 2019), and overall language competency. One particularly fruitful area of L2 receptive research has explored the relationship between lexical complexity and L2 text difficulty (Brunfaut & Révész, 2015; Révész & Brunfaut, 2013). To date, comparatively little research has addressed the issue of how lexical complexity contributes to written and spoken text difficulty in L2, and, particularly, scant attention has been paid to the investigation of lexical complexity in videotext. Nonetheless, the studies that do exist have identified some aspects of lexis that seem to influence comprehension of spoken texts (Brunfaut & Révész, 2015; Matthews & Cheng, 2015; Révész & Brunfaut, 2013; Stæhr, 2009).

The contemporary view of lexical complexity—also labeled as lexical richness (see, e.g., Read, 2000)—asserts that it is a multidimensional, multifaceted construct that generally encompasses three interconnected facets; lexical sophistication, lexical density, and lexical diversity (Lu, 2012). Various metrics and measures have been proposed to tap into these three dimensions of lexical complexity. Many of these metrics have been extensively examined in second language studies and were found to be reliable and valid indices for both receptive and productive language skills. However, to our knowledge, no previous research has explored to what extent these indices can be used as reliable and valid measures of lexical complexity in L2 videotext difficulty. Because visual information plays a critical role in language processing and understanding (Ferreira, Foucart, & Engelhardt, 2013; Knoeferle & Guerra, 2016), it is reasonable to suggest that processing and recognizing words in videotexts is not the same as in audio-only texts.

##### 3.4.1.2.1 Lexical Sophistication

Lexical sophistication or rareness generally refers to the proportion of sophisticated or

difficult words in text (Laufer & Nation, 1995). While the definition of sophisticated or difficult words has yet to be agreed upon (Daller, Milton, & Treffers-Daller, 2007), sophisticated lexis has been associated with infrequent or uncommon words. This conceptualization is motivated by the assumption that less frequent words in natural language are less likely to be known, hence not recognized by language learners (Ellis, 2002; Gries & Ellis, 2015). Beyond frequency, there has been little consensus among researchers on how lexical sophistication should be defined (Kim, Crossley, & Kyle, 2018; Milton, 2009). For instance, sophisticated words have been identified as those that are less concrete and less imageable (Saito, Webb, Trofimovich, & Isaacs, 2016; Salsbury, Crossley, & McNamara, 2011), words that are acquired at a later age, words that are widely used in academic contexts (Coxhead, 2000), words that are contextually less diverse (McDonald & Shillcock, 2009), and words that elicit slower response times in lexical decision and naming tasks (Balota et al., 2007).

In the sentence processing literature, word difficulty has been linked to word surprisal. According to the *surprisal theory* (Hale, 2001; Levy, 2008), the processing difficulty incurred by a word is inversely proportional to its expectancy. That is, the less expected a word is in a given context, the higher its surprisal, and hence the greater its processing demand.

Beyond counting the frequency of sophisticated words, recent investigations have expanded the scope of the construct by considering various other lexical features, including, but not limited to, hypernymy, polysemy, and psycholinguistic word norms (Kim et al., 2018; Kyle & Crossley, 2016). Findings of this line of research have generally demonstrated that certain aspects of lexical sophistication tend to aggregate and overlap among each other, suggesting that they form consequently new (latent) variables. For instance, Kim et al. (2018) explored if a wide set of lexical sophistication indices (e.g., concreteness, orthographic density, hypernymy, and n-gram frequency and association strength) can be aggregated into co-occurring features (components) and whether these microfeatures were predictive of language abilities (i.e., L2 writing proficiency and L2 lexical proficiency) as well as indicative of longitudinal lexical growth. The results of the principal components analysis showed a strong overlap or interconnectivity among several indices of lexical sophistication forming as such several dimensions or components. These components explained a considerable amount of variance in L2 lexical and writing proficiency and revealed developmental trends in L2 beginning learners' use of lexis over one year. The findings of this study suggest that lexical sophistication is more likely a multidimensional, multifaceted construct, that goes beyond lexical frequency. In the sections

that follow, I will review some lexical proprieties that have been proposed to contribute to lexical sophistication.

**Lexical Frequency**

Corpus-based lexical frequency is the hallmark measurement of lexical sophistication. Frequency analysis is grounded on Zipfian law which posits that a small number of word types occurs very frequently in a language and makes up the majority of running words in text (Zipf, 1945). This phenomenon has been the locus of extensive research in corpus linguistics and vocabulary studies and a large body of empirical evidence confirmed its critical role on text comprehension, vocabulary learning, and language development. In essence, it is believed that words occurring more frequently in natural language are learned earlier, processed more quickly, and retrieved with greater accuracy than low-frequency words (Balota & Chumbley, 1984; Kirsner, 1994). Because language learners are regularly exposed to more frequent words than less frequent words, they are expected to acquire the most frequent words first and as they progress in their language development they learn less frequent words later (Muljani, Koda, & Moates, 1998)

Lexical frequency has almost always been operationalized in reference to large standard corpora of written and spoken texts such as the 100-million word British National Corpus (BNC: Leech & Rayson, 2014) and the 450-million word Corpus of Contemporary American English (COCA: Davies, 2010). To estimate word frequency, L2 researchers have generally used two main approaches, namely a band-based and count-based frequency approach (Crossley, Cobb, & McNamara, 2013). In the band-based approach, frequency bands are developed in multiple steps. First, word families (inflections and obvious derivations) are given a frequency rating based on the sum of all of the members of the word family and then a cutoff threshold is decided to group words into lists of word families. These word families are then refined based on word distribution or range and genre considerations (spoken and written). Finally, word families are assembled into bands, so that a text can be said to have 80% of words belonging to the K1 most frequent words. Measures based on the former approach often involve calculating the percentage or proportion of words that appear in the frequency bands derived from large corpora (BNC: Leech & Rayson, 2014) and (CELEX: Baayen, Piepenbrock, & Gulikers, 1995). Examples of these indices include Lexical Frequency Profiles (Laufer & Nation, 1995) and P_Lex (Meara & Bell, 2001). On the other hand, the frequency-based approach (Tuldava, 1996) often involves averaging all individual word frequencies in a text to a one-integer overall frequency rating of the text. To illustrate this point using an example provided by Crossley et al. (2013), the frequency value of the sentence "the cat sat on the mat" based on

the BNC frequency norms would be as follows: *The* (6,041,234) *cat* (3844) *sat* (11,038) *on* (729,518) *the* (6,041,234) *mat* (569) and the average word frequency value for the sentence is 2,137,906.17 (*SD* = 3,036,494.47).

In terms of their utilities, measures based on the two approaches have their advantages and disadvantages. For example, one apparent advantage of the band-based approach is the sense of clarity and interpretability that come with a grouping that can be of use to language teachers, course designers, and researchers. However, grouping words onto bands of word families lead to fewer distinctions among words, and hence nuance details may not be captured (Crossley et al., 2013). According to Milton (2009), "there is no absolute rule as to what point at which a word stops being frequent and become infrequent" (p.131). On the other hand, the count-based approach does not assume that words should be arbitrarily divided into groups or bands of 1000 word families. It relies on large reference corpora to calculate individual word frequencies. Previous work showed some evidence that the use of count-based frequency indices for capturing finer details of lexical sophistication (Crossley, Salsbury, McNamara, & Jarvis, 2011; Kyle & Crossley, 2015). A key limitation of this approach, however, is the lack of interpretability which makes gaining insights to inform pedagogical practices more difficult (Crossley et al., 2013; Crossley, Salsbury, et al., 2011).

The chosen approach notwithstanding, the word frequency effect is well-promulgated and its impact on text comprehension has been empirically confirmed. Given that and the instrumental role of word knowledge on text comprehension, it is reasonable to suggest that text with a greater proportion of less frequent lexis poses a greater processing challenge for language learners. However, previous research on L2 spoken text difficulty has generally brought inconclusive findings. For example, Nissan, DeVincenzi, and Tang (1995) found that TOEFL test takers experienced greater difficulty comprehending dialogue items that contain infrequent words (i.e., a word not included in Berger's list (1977) of 100,000 common words). While also using Berger's list, Kostin (2004) found no significant impact of infrequent words on listening difficulty, but her findings have also demonstrated that dialogues tended to be more difficult when comprehension of less frequent vocabulary was required to respond correctly.

In a more comprehensive study of listening task difficulty, Révész and Brunfaut (2013) investigated whether speech rate, linguistic complexity (e.g., phonological, lexical, syntactic, and discourse complexity), and explicitness of spoken texts influence advanced language learners' comprehension performance. For assessing lexical frequency in their study, Révész and Brunfaut (2013) utilized West's General Service List of English Words

to calculate the percentage of all words, function words, and content words appearing in the list's first (K1) and second thousand (K2) most frequent English word families. The findings of regression analysis demonstrated that the proportion of K1 function words and lexical density had a strong effect on listening difficulty, accounting for 40% of the variability in listening comprehension. In contrast, other studies failed to observe a relationship between lexical frequency and spoken text difficulty (Yanagawa & Green, 2008; Ying-hui, 2006).

**N-gram Frequency and Formulaic Expression**

It is understood that multiple lexical items act as a single item and they aid listeners (Wray, 2000). In addition to counting the frequency of individual lexical items in a text, L2 researchers have also explored whether the presence of formulaic expressions—multiword units whose meanings are distinct from those of their individual parts—contributes to listening difficulty (Brunfaut & Révész, 2015; Révész & Brunfaut, 2013; Ying-hui, 2006). The overall finding of these studies demonstrated that the presence of formulaic expressions in spoken texts has no significant impact on L2 listening difficulty. This finding is unexpected and runs counter previous studies (Ellis, Simpson-vlach, & Maynard, 2008; Kostin, 2004), in which formula expressions appeared to increase listening demands.

Alongside issues related to selecting an appropriate word list, the majority of listening research has generally regarded lexical frequency as a dichotomous factor (i.e. whether a text has low-frequency words or not) and failed to account for the quantity of low-frequency words and their nature across spoken texts, with an exception of Brunfaut and Révész (2015) and Révész and Brunfaut (2013) who also counted the frequency of low-frequent words in their investigations.

Because frequency measures are based on how many times a word or word family occurs in a corpus as a whole, they might not accurately capture how widely a particular word or word family is used in the corpus (Kyle & Crossley, 2015). Alternatively, measures of word range, also known as dispersion and contextual diversity, are often used (Gries, 2008), usually by counting in how many documents in the corpus a word occurs. There seems to be no previous listening research that has explicated range indices and how they might contribute to our understanding of spoken or video texts.

**Psycholinguistic Properties of Word**

Cognitive scientists and psycholinguists have long been interested in exploring what properties of the word affect their processing and learnability (Grainger, 1990; Whaley, 1978). Among several attributes of words, the psycholinguistic proprieties of *concreteness,*

*meaningfulness, imageability, word familiarity*, and *age of acquisition* have been investigated in text complexity and readability literature. These word attributes or norms have been empirically found to interact with lexical processing speed and accuracy. *Concreteness* refers to the extent to which content words in a text refer to concrete objects/events or abstract concepts/ideas. A word can be said to be concrete, if one can simply point to the object it signifies (e.g., a chair or an apple), while if a word can be only described by other words (e.g., happiness), it can be considered more abstract (Brysbaert, Warriner, & Kuperman, 2014). There is ample psycholinguistic evidence that shows processing abstract words is more challenging than concept words (Paivio, Walsh, & Bons, 1994). *Meaningfulness* refers to how a word is closely related to other words (Toglia & Battig, 1978). Word *imageability* refers to how easy is to create a mental image of the word. Word familiarly is closely related to the lexical frequency and it describes how commonly a word is experienced. *Age of acquisition* score is based on human judgments of the age at which a particular word is learned (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). The other measures are based on adults' ratings of content words on 7-point scales, with higher scores reflecting easier processing. Perhaps the most common repository for linguistic and psycholinguistic information on English word attributes is the Medical Research Council (MRC) psycholinguistic database (Coltheart, 1981; Wilson, 1988). MRC database includes scores of subjective human judgments on 26 different linguistic and psycholinguistic attributes of as many as 150,837 English words.

#### 3.4.1.2.2 Lexical Density

Originally coined by Ure (1971b), lexical density refers to the proportion of content words (nouns, verbs, adjectives, and sometimes adverbs) to the total number of words in a text. Content words, as opposed to function words (e.g., preposition, interjections, pronouns, conjunctions and count words), carry more information in language. Accordingly, lexical density is generally regarded as an index of information packaging (Johansson, 2009). Processing texts with high lexical density presumably exerts a greater cognitive load on L2 listeners (Bloomfield et al., 2010). Lexical density varies across text types and registers. For example, when compared to written texts, spoken texts reportedly have a lower lexical density (Halliday & Hasan, 1985), with the majority of spoken texts having a lexical density of under 40% (Ure, 1971b). How lexical items are defined plays an important role in measuring lexical density.

To date, few studies have examined the role of lexical density on L2 listening comprehension difficulty. For example, Buck and Tatsuoka (1998) found that listening difficulty

tends to increase when a greater percentage of content words surrounded the information relevant to task completion. Similarly, Révész and Brunfaut (2013) found that lexical density was a key predictor of listening task difficulty, explaining 40% of the variance in language learners' performance in a listening comprehension test. But in a later study, Brunfaut and Révész (2015) observed that lexical density had an insignificant correlation with listening task difficulty.

#### 3.4.1.2.3 Lexical Diversity

Lexical diversity is often described as the range and variety of vocabulary deployed in a text. Often used interchangeably with the terms lexical richness (Tweedie & Baayen, 1998), vocabulary richness (Hoover, 2003), and lexical variation (Read, 2000), measures of lexical diversity are concerned with how many unique words are in a text, with higher diversity being associated with a more varied use of lexis. Arguably, a text that contains a wider diversity of unique words is more difficult because language learners need to decode a large number of unique words in a limited time (Bloomfield et al., 2010). Previous studies have found lexical diversity to be a significant predictor of spoken text difficulty (Révész & Brunfaut, 2013; Rupp, Garcia, & Jamieson, 2001). In their study, Rupp et al. (2001), for example, found listening item difficulty was commensurate with lexical diversity, as measured by TTR. Révész and Brunfaut (2013), using vocd-D, found a significant association between lexical diversity and L2 listening success.

Traditionally, lexical diversity is assessed by calculating the ratio of different words (types) to the total number of words (tokens), the so-called type-token ratio, or TTR (Chotlos, 1944; Templin, 1957). One key problem of TRR and indeed all indices based on word frequency is that their values are affected by text length; that is as the number of tokens increases the likelihood of those tokens being unique becomes very low (H. Heaps, 1978). To address this issue, researchers have attempted some mathematical transformations on TTR, for instance, Mean Segmental TTR, Root TTR, and Corrected TTR, but these variants seem not to overcome the problem entirely.

More recently, several alternative measures of lexical diversity have been proposed in recent years such as D (Malvern & Richards, 1997; Malvern, Richards, Chipere, & Durán, 2004), Measure of Lexical Textual Diversity (MLTD; McCarthy, 2005; McCarthy & Jarvis, 2010), and the Moving-Average Type-Token Ratio (MATTR; Covington & McFall, 2010), each purport to measure lexical diversity, while having little or no effect of text length. The D-value is derived, often by computer software such as CLAN (MacWhinney, 2000), through a series of computations of the TTR on samples of different text lengths (typically

ranging from 35 to 50 tokens) after which a random sampling TTR curve is computed (but see McCarthy and Jarvis 2 007, for a critical appraisal of D). McCarthy and Jarvis (2007) provided theoretical and empirical proof that void-D–specifically its sampling and curve-fitting procedures–has shortcomings and proposed HD-D as a better alternative. HD-D is similar to D but based on the hypergeometric distribution function (Wu, 1993). MLTD splitting a text into segments that have a type-token ratio (TTR) value of .72 (see, McCarthy & Jarvis, 2010, for more details). The total segments in the text are then divided by the total number of words in the text (the forward value). The whole process is then repeated starting at the end of the text rather than the beginning (the reverse value). The forward and reverse values are then summed and divided by 2 for the MLTD value of the text. MLTD was used in Crossley, Salsbury, and McNamara (2009) and later validated in (McCarthy & Jarvis, 2010)

While the newer indices implement sophisticated techniques, Treffers-Daller, Parslow, and Williams (2018) found traditional measures of lexical diversity (e.g., TTR and Root TTR) more effective in discriminating between texts of different CEFR levels than more sophisticated measures (e.g., D, HD-D, and MLTD), provided text length is kept constant across texts. Besides what approach yields the most accurate account of lexical diversity, researchers have also debated what constitutes a different word or type (i.e., whether all vs. some inflected forms should be considered as tokens of the same type). As shown in Durán, Malvern, Richards, and Chipere (2004) and Treffers-Daller (2013) and later in Treffers-Daller et al. (2018), different lemmatization approaches may generate different lexical diversity values for the same texts. In the context of L2 research, the decision of whether to compute indices of lexical diversity based on the word, lemma, or word family depends on the objectives of the research. That is when the focus is on measuring the learner's productive ability (e.g., speaking and writing), lemmatization is probably not the best option as critical information of the learner's vocabulary knowledge would be lost. But when the objective of the research is on learner's receptive ability then using any of three measurement units seems reasonable, yet further research is needed to explore which unit is more effective.

It should be noted that it is unclear how word type is operationalized in previous research on spoken text difficulty (e.g., Brunfaut & Révész, 2015; Révész & Brunfaut, 2013) which makes comparing findings and evaluating approaches difficult. Another shortcoming of previous work is the use of an unreliable or single measure of lexical diversity. Because lexical diversity indices gauge lexical diversity in different ways, using multiple indices is warranted to capture all aspects of lexical diversity (McCarthy & Jarvis, 2010).

TABLE 3.3: Lexical Complexity, their Impact & Supporting Empirical Evidence

| | **Hypothesized Impact** | **Empirical Evidence** |
|---|---|---|
| Lexical frequency | High-frequency words elicit shorter processing and fixation times | (Kuperman et al., 2012) |
| Lexical density | A text that has more content words (i.e. noun, verbs, adjectives, and some adverbs) are more difficult to understand because content words carry more information | (Brunfaut & Révész, 2015; Buck & Tatsuoka, 1998; Révész & Brunfaut, 2013) |
| Lexical diversity | The diversity of unique words in texts leads to more difficulty because comprehender needs to decode a large number of unique words | (Révész & Brunfaut, 2013; Rupp et al., 2001) |
| Concreteness | A word that refers to an abstract concept or idea is more challenging than a word that refers to a more concert object. | (Barber, Otten, Kousta, & Vigliocco, 2013; Paivio et al., 1994; Schwanenflugel, 1991) |
| Age of acquisition | Early-learned words have an advantage over late AoA words in lexical processing speed and accuracy. | (R. M. DeKeyser, 2013; Saito, 2013) |
| Familiarity | Familiar words are processed and recognized faster than less familiar words. | (Connine, Mullennix, Shernoff, & Yelen, 1990) |
| Meaningfulness | How a word is related to other words in meaning. Semantically related words, especially nouns, are acquired more easily in first languages than in second languages. This is because L2 learners tend to have fewer and more varied word associations than do L1 speakers. | (Balota & Chumbley, 1984; Zareva, 2007) |
| Imagability | A high-imagery word is more likely to be recalled because it evokes a mental image quickly and easily. | (Strain & Herdman, 1999; Strain, Patterson, & Seidenberg, 1995) |

To recap thus far, lexical complexity is generally assumed to be a componential construct that conflates a set of subordinate concepts such as lexical sophistication, density, and diversity. In the last few years, researchers have begun to interrogate the relationship between lexical complexity and spoken text difficulty. The findings of this research strand underscore the role of lexical complexity on spoken language understanding. However, to date, we know little about the interaction between lexical complexity and videotext difficulty. Given the audiovisual nature of videotext, it is very plausible to hypothesize that recognizing and processing lexis in videotext is different and the impact of lexical complexity on videotext may not be the same as on spoken or written texts. However, such a claim warrants empirical investigation.

#### 3.4.1.3 Syntactic Complexity

Syntactic processing is an essential component in L2 spoken text comprehension process (Rost, 2013). Therefore, it is reasonable to assume that texts with greater syntactic complexity are associated with increased listening difficulty (Révész & Brunfaut, 2013). Syntactic complexity generally refers to the variety and sophistication of syntactic structures. Traditionally, syntactic complexity has been measured using a simple measure, the mean sentence length. However, it is now believed that syntactic complexity is a multidimensional construct with at least three subdimensions: complexity by subordination, phrasal complexity, and overall complexity (Norris & Ortega, 2009). Over the last decade, numerous indices and methods have been proposed to tap into these different dimensions of syntactic complexity. In a recent review of syntactic complexity indices, Norris and Ortega (2009) warranted that some of the syntactic indices used in previous research are redundant and measure the same underlying construct and as such, they recommended researchers to select their indices carefully.

The bulk of previous L2 research on syntactic complexity has been predominantly concerned with the products of language learners, for example, to investigate how syntactic structures become more complex or advance in language learner speech and writing as they progress in their learning, and very few studies have examined whether different syntactic structures affect L2 listening text comprehension and difficulty (Blau, 1990; Brunfaut & Révész, 2015; Kostin, 2004; Révész & Brunfaut, 2013). The findings of previous studies were mixed. For example, Cervantes and Gainer (1992) found a relationship between lower-degree subordination and comprehension of short lectures. In their study on text readability assessment, Vajjala and Meurers (2012) examined the contribution of a

variety of linguistic complexity features and found 14 syntactic complexity features alone achieved an accuracy of 71.2% for predicting the grade levels of reading texts. However, the findings of other studies demonstrate no significant effect of any of the three sub-constructs of structural complexity on listening difficulty—complexity by subordination (Blau, 1990; Brunfaut & Révész, 2015; Kostin, 2004; Ying-hui, 2006), phrasal complexity (Révész & Brunfaut, 2013), and overall complexity (Révész & Brunfaut, 2013; Ying-hui, 2006).

Recently, Jin, Lu, and Ni (2020) conducted a large-scale study to assess the quantitative differences in syntactic complexity in 3,368 teaching texts in use in teaching 12 EFL grade levels (primary school Grades 1–6, middle school Grades 1–3, and high school Grades 1–3) in China. The researchers found that 8 syntactic complexity indices had significant between-level differences with moderate to large effect sizes and 5 indices were significant predictors for classifying grade levels.

To sum up, previous studies have brought mixed findings regarding the impact of syntactic complexity on spoken text difficulty, and to the best of my knowledge, there has been no systematic effort to investigate the contribution of syntactic complexity to videotext difficulty.

#### 3.4.1.4 Discourse Complexity

In a typical speech, speakers rarely organize complex thoughts into single utterances; instead, they develop their thoughts and expand upon them over many utterances in the discourse. The speaker's narrative, however, should be cohesive so that it becomes more comprehensible for the listener. Cohesion is crucial for comprehension and it generally refers to the explicit cues in text that assist readers and listeners in connecting ideas within the text (Gernsbacher, 1990; Graesser, McNamara, & Louwerse, 2003; Halliday & Hasan, 1976). Examples of explicit cohesion cues are connectives (e.g., because, therefore, and so), reference (e.g., pronouns), and overlapping words in the text. These cohesive cues signify to the reader or listener that the same or similar concepts are being referred to across consecutive sentences or paragraphs (Crossley, Kyle, & McNamara, 2016).

Discourse cohesion is believed to be involved in the determination of text ease or difficulty (McNamara et al., 1996; McNamara & Kintsch, 1996b; Sheehan et al., 2010). Texts with very few cohesive cues are assumed to be more difficult because text readers or listeners need to invoke more mental resources to fill in gaps in the text (Zwaan & Radvansky,

1998). But when there are few or no gaps, then a text is seamlessly moving from one point of information to another while giving the listener all the help he or she needs to build new knowledge. Research in L1 text comprehension has shown that text cohesion is more beneficial to low-knowledge readers (McNamara & Kintsch, 1996b) who rely on cohesion cues to fill in the gaps in their background knowledge. In contrast, for high-knowledge readers, the absence of local cohesion cues in a text seems to be more helpful because it prompts the generation of inferences that connect ideas in the text and to the reader's prior knowledge, thus enhancing text comprehension (McNamara et al., 1996; McNamara & Kintsch, 1996b).

Two types of text cohesion have been frequently explored in the literature: referential cohesion and causal cohesion. Referential cohesion refers to "the overlap in words or semantic references, between units in the text such as clauses, sentences, paragraphs" (McNamara et al., 2014, p. 22), whereas causal cohesion refers to the degree to which causal relationships are explicitly stated in a text, for example, using connectives such as *because, therefore*, and *consequently* (McNamara et al., 2014).

While cohesion and coherence relations have been examined extensively in L2 writing, a few studies to date have explored the relationship between cohesion and listening difficulty, and they have yielded contradictory findings. For instance,Ying-hui (2006) investigated L2 listening difficultly as the function of cohesion, which she operationalized as the extent to which the elements of the opening sentence were represented in the remainder of a passage. Ying-hui (2006) then obtained expert judgments on text cohesion and found that a significant correlation between easier listening items received higher cohesion ratings. Révész and Brunfaut (2013) analyzed cohesion in terms of different types of connectives, causal, intentional, temporal, and spatial cohesion, but only causal content emerged as a significant predictor of L2 listening difficulty. In Brunfaut and Révész (2015), texts with higher referential cohesion, as measured by (adjacent) argument overlap and stem overlap shared across utterances in the listening passage, are significantly associated with easier listening tasks. Nissan et al. (1995), however, found no effects for cohesion in exploring the relationship between item difficulty and whether passages contained explicit lexical links (e.g., repetition) or structural links (e.g., anaphora) between the speakers' utterances.

Narrativity is another characteristic of spoken discourse that is believed to facilitate comprehension. A narrative text tells a story and is closely affiliated with conversational speech. For most L2 listeners, a narrative text is more comprehensible than any other

TABLE 3.5: Syntactic and discourse Complexity, their Impact & Supporting Empirical Evidence

| | **Hypothesized Effect** | **Empirical Evidence** |
|---|---|---|
| Syntactic complexity | Processing long and complex sentence structures are more taxing. | (Cervantes & Gainer, 1992; Gibson, 1998) |
| Narrativity | Narrative text tells a story and is closely affiliated with conversational speech. Non-narrative texts on a less familiar topic are presumably more complex than narrative texts. | (McNamara et al., 2014) |
| Discourse cohesion | Low cohesive text is typically more difficult to process because readers or listeners need to fill in gaps in the text when making inferences | (Brunfaut & Révész, 2015; Graesser et al., 2003; McNamara & Kintsch, 1996b; Révész & Brunfaut, 2013; Ying-hui, 2006). |

type of text (Rubin, 1994). On the other hand, non-narrative texts on less familiar topics are presumably more difficult. Typically, narrativity is determined through analysis of lexical familiarity, syntactic complexity, and grammatical features, such as length of noun phrases, the relevance of pronouns, and the number of adverbs (McNamara et al., 2014).

### 3.4.2 Summary of Verbal Complexity

Thus far, I have reviewed studies on verbal complexity from different research areas and outlined key attributes of verbal language that have been empirically found to impair (spoken) text understanding. While language is a key component of videotext, it is certainly not the only one. Videotexts are a multimodal type of text, and our investigation of videotext difficulty cannot be said to be complete without interrogating the contribution of visual complexity to language learners' perception of difficulty in L2 videotexts. In the next section, I address the question of how visual imagery may impact videotext understanding.

### 3.4.3 Visual Complexity

Research on visual complexity has been carried out in many areas, including cognitive science, marketing, psychology, human-computer interaction, aesthetics, and psychobiology (e.g., Braun, Amirshahi, Denzler, & Redies, 2013; Pieters, Wedel, & Batra, 2010; Tuch, Bargas-Avila, Opwis, & Wilhelm, 2009). That being said, the notion of visual complexity remains elusive and hard to pin down, however. Originally proposed for static images, visual complexity broadly refers to the amount of detail or intricacy contained within an image (Snodgrass & Vanderwart, 1980). Perceiving a visual stimulus more or less complex is also suggested to be influenced by several other factors, including, for example, the type and quantify of elements it contains, their spatial distribution or layout, variety of colors, and so forth (Palumbo, Makin, & Bertamini, 2014). C. Heaps and Handel (1999) contended that visual complexity corresponds to the degree of difficulty people encounter when describing a visual stimulus. Building upon different conceptualizations of complexity, numerous approaches and techniques have been proposed to gauge complexity in visual stimuli (see Donderi, 2006, for comprehensive review). Consequently, "any study of complexity faces the problem of selecting a definition and a measure of visual complexity" (Palumbo et al., 2014, p.3).

For assessing visual complexity, researchers have utilized both qualitative and quantitative approaches. A qualitative approach often involves human participants rating visual complexity using a predefined subjective scale. Despite a high correlation with human perception, such subjective rating scores are costly to obtain, less consistent, and cannot be generalized to other images. On the other hand, quantitative measures seek to provide a reliable and consistent numerical value that corresponds to some behavioral data of complexity in visual stimuli.

In psychological and behavioral research, complexity in visual stimuli is often determined using reaction times and human subjective ratings. It is assumed that the perception of visual complexity is subjective and individuals as such tend to rate visual complexity differently (e.g., Corchs, Ciocca, Bricolo, & Gasparini, 2016). In the fields of computer vision, image processing, and human-computer interaction, the focus has been primarily on devising reliable objective measures of complexity, for example, by relying on spatial and statistical proprieties of visual stimuli that correlate with human perception of complexity. A popular method for computing visual complexity is image compression. Specifically, it is assumed that complexity and compression are intimately related concepts. That is, an image of complex visual stimuli requires a larger file size when compressed than

an image of simple stimuli (Salomon, 2004; Sayood, 2017). Using this rationale, several studies have found a correlation between compressed image sizes and human perception of visual complexity (e.g., Donderi & McFadden, 2005).

With the vast majority of visual complexity research devoted to static images, we know little about visual complexity in dynamic multimedia such as videotexts. To date, there is understandably a dearth of research on what visual complexity actually entails in videotexts. The dynamic and multimodal nature of videotext makes the task of underpinning what visual complexity entails in videos is more challenging.

Computational measures of visual complexity should account for both spatial and temporal dimensions. Unlike its constrained use in the literature to exclusively refer to the image complexity, in this thesis, the construct of "visual complexity" is very broad and it conflates any visual and spatiotemporal attributes of video that makes videotext processing more effortful and challenging.

A useful departure point for understanding video visual complexity is research on visual attention. This is because the human brain is unable to simultaneously process everything that is there in the environment and attention plays a fundamental role in regulating and prioritizing which part of the multitude of sensory data is currently worthy of processing. One definition of visual attention is that it is "the process by which some objects or locations are selected to receive more processing than others" (Findlay & Gilchrist, 2003, p. 35). When we pay attention [3] to an object, we deploy more resources for processing it, resulting in a rich representation of the object in perception (W. James, 2007). Further, this selection permits the reduction of complexity and information overload (Evans et al., 2011). Last but not least, attention plays a key role in object recognition. Since the object recognition mechanisms cannot process all objects in the visual field at once, visual attention helps in selecting digestible input for recognition (Evans et al., 2011). For this reason, representations of attended objects vary from those of unattended objects (DeSchepper & Treisman, 1996).

Admittedly, the topic of visual attention is vast and could not be duly reviewed in a single section but interested readers can refer to some books on this topic (e.g., Johnson & Proctor, 2004; Pashler, 1999). Here, what we are mainly interested in is how attention makes coordinating limited cognitive resources possible for attending to and processing some stimulus but not others, and more importantly what properties of visual stimulus

[3] The term attention can refer to other psychological phenomena such as remaining alert for long periods, here the term refers exclusively to perceptual selectivity.

make them more worthy of attention and hence of processing. After discussing the role of attention, I then try to garner pieces of empirical evidence for the different attributes of visual imagery that seem to impede the understanding of dynamic multimedia. Throughout this section, I use the term "video visual complexity" as an umbrella term to refer to all these visual attributes, being static or dynamic ones.

#### 3.4.3.1 Visual Attention

A typical visual scene is complex and replete with a vast amount of information. Nevertheless, we experience a seemingly effortless understanding of our visual world. This is because at any given moment the vision faculty in our brains rapidly prioritize and select the information that is relevant to our task and disregard those that are irrelevant. This attentional mechanism is commonly known as *selective attention* and it does not apply to visual input per se but also applies to all types of information coming to our senses; for example, the cocktail-party effect is a well-known exemplar of selective auditory attention.

Visual attention has been metaphorically compared to a "spotlight" that only selects parts of the visual field making them more prominent, or more salient, while simultaneously, suppressing irrelevant objects and making them more or less invisible (Posner, Snyder, & Davidson, 1980). Instead of a single cognitive process, visual attention is generally thought of as a family of processing resources or cognitive operations that mediate the selection of relevant and the filtering out of irrelevant information from cluttered visual scenes (Evans et al., 2011). Although visual attention is closely synchronized with where the eyes fixate, it is also possible to be aware of objects in the peripheral region, outside the central foveal vision. In visual attention literature, these two attentional mechanisms respectively describe overt attention and covert attention. More concretely, overt attention is a shift of attention to a relevant location that coincides with the movements of eyes or head, whereas a covert shift of attention requires no such deliberate movements (Carrasco, 2011). An example of covert attention is car driving, where a driver keeps her eyes on the road while simultaneously covertly monitoring the status of signs and lights. While attending overtly to a visual stimulus is crucial, the inhibition of irrelevant information is also as important for it reduces complexity and information overload (Evans et al., 2011).

#### 3.4.3.2 Factors Affecting Visual Attention

Visual attention, as alluded to earlier, is a selective process, but what determines what we select and process first in a complex scene? Research on attention selection has traditionally proposed two major groups of factors that drive attention allocation; bottom-up or stimuli-driven factors and top-down or user-driven factors (Evans et al., 2011; Wolfe & Horowitz, 2017). Bottom-up factors refer to the preattentive attributes [4] of an object that attracts observer's attention regardless of his/her desire and intention (Theeuwes, 1992, 2010). The size, color, shape, and orientation of the visual stimulus are among those attributes. Object saliency, which will be discussed later, is assumed to drive attention. On the other hand, top-down attention is determined by cognitive phenomena such as prior knowledge, experience, current goal, reward, and so forth (see Wolfe & Horowitz, 2017, for a review). A prototypical example of top-down attention is searching for a specific object, for example, a car key amongst different objects on a disk.

Each supported by a wealth of empirical studies, the two theories provide a cogent account for visual guidance, which has resulted in a theoretical stalemate. Recently, some researchers have put forth a view that integrates bottom-up and top-down factors (Gaspelin & Luck, 2018, 2019).

A typical real-world context is dynamic and it is expected a combination of both bottom-up and top-down factors modulate visual attention. Much recent research has focused on the interplay between the two types of factors and the finding of this work has been that visual attention is not simply controlled by either bottom-up or top-down factors but rather it seems to be modulated by the two (Evans et al., 2011; Gaspelin & Luck, 2019). For instance, the *Guided Search model* (Wolfe, 1994a, 2007) posits that the deployment of attention is determined by some weighted combination of bottom-up and top-down factors.

Another contested issue is how visual attention is deployed to objects in the visual field. For several decades, there has been some agreement that there are essentially two independent stages of visual processing (e.h., Broadbent, 1958; Treisman & Gelade, 1980). An early preattentive stage operating in parallel across the visual field and a later stage that can only handle one (or a few items) at the same time (Theeuwes, 2010). This dichotomy has recently been under scrutiny and researchers now believe the boundary between the

[4] Wolfe and Utochkin define a preattentive feature as "a feature that guides attention in visual search and that cannot be decomposed into simpler features " (2019, p. 19).

two stages is not independently exclusive as originally assumed. Theeuwes (2010) suggested that when light initially hits the retina, visual attention is completely driven by the visual proprieties of the stimuli, and only later visual attention, through massive recurrent feedback processing, will be driven by the goals, intentions, and expectations of the observer in a top-down fashion.

Recently, Wolfe and Horowitz (2017) provided five key factors that guide people's attention to an object in a visual scene; bottom-up salience, top-down feature guidance, scene structure and meaning, the previous history of search, and the relative value of the targets and distractors. In the sections that follow I disuses some visual attributes that interfere with visual attention.

#### 3.4.3.3 Visual Attention and Object Saliency

Generally, attention is naturally attracted to salient objects in the visual field (Wolfe & Horowitz, 2017). When viewing a natural scene without a specific task in mind (as in free watching), object saliency is more likely to guide human visual attention. But what makes an object salient? Saliency is the quality of the objects that make them pop out. A distinctly different object, for example, in color, shape, and orientation, can stand out from its neighboring objects. Because at any given moment visual information is rich and human attentional capacity is severely limited, silent objects are more likely to be processed and recognized more rapidly while other objects are simply ignored. However, when a visual search is guided by a task, observers may inhibit salient distractors and prioritize features of objects that match the features of the search target (Gaspelin & Luck, 2019; Vatterott & Vecera, 2012). For example, when searching for a strawberry among other fruits, we might restrict our attention to red items only and ignore anything else.

While it is generally assumed that a salient object has an inherent power to involuntarily attract attention, regardless of the observer's immediate task or intention, recent research has suggested that observers learn to suppress salient distractors (Vatterott & Vecera, 2012). That is, visual attention may also be guided away from objects that mismatch anticipated characteristics of the target object (Gaspelin & Luck, 2019). An account of this inhibitory mechanism is postulated by the *signal suppression hypothesis* (see Gaspelin & Luck, 2018, for a review). Converging evidence for the existence of such suppression mechanism has begun to emerge in recent research employing more advanced

techniques based on psychophysics, eye-tracking, and event-related potentials (De Tommaso & Turatto, 2019).

In addition to perceptual studies, findings from research in multimedia learning have also accentuated the critical role of visual attention in multimedia learning in light of the limited capacity of human working memory (Mayer, 2009). It is generally recommended that multimedia learning materials should be designed in a way that allows learners to pay more attention to the most important parts of learning materials, and less attention to those that are irrelevant. A well-established and empirically tested design technique is the *singling principle*. The singling principle (also known as the attention cuing principle) is based on the finding that people learn better from materials that have cues highlighting relevant information (see Schneider, Beege, Nebel, & Rey, 2018, for a review). Common cuing techniques in dynamic multimedia include changing the text's font and style, increasing the luminance of specific objects of the visual display, and using arrows to point learners to a specific region (De Koning, Tabbers, Rikers, & Paas, 2009). In a recent meta-analysis study in which 103 studies on singling effect were examined, Schneider et al. (2018) reported a positive effect of singling in learning retention and performance. They also found the incorporation of singling in multimedia materials reduces cognitive load significantly.

Computational models of object saliency detection simulate the human visual system to perceive saliency in both static and dynamic pictures (Borji & Itti, 2013). When not guided by a task, object saliency can be a good predictor of gaze movement (Coco, Malcolm, & Keller, 2014).

These object proprieties and others are commonly used in the development of computational saliency models that predict what object in a display is more likely to attract attention (Koehler, Guo, Zhang, & Eckstein, 2014). The accuracy of these models has improved over the years and current state-of-the-art models do reasonably well. However, as noted earlier, bottom-up saliency is not the only factor that guides attention. Attention is also modulated by the observer's intention, goal, and the task at hand.

Inspired by theoretical works such as *Feature Integration Theory* (Treisman & Gelade, 1980) and *Guided Search Model* (Wolfe, 1994a), stimulus-driven computational models determine visual saliency based on spatial proprieties of the image such as the difference between physical characteristics of the image, including differences or contrasts in color, intensity, and orientation, relative to its background. These models—also known as saliency maps—emphasize the role of bottom-up process based on low-level features

of the image over top-down process (Borji & Itti, 2013; Q. Zhao & Koch, 2013, for review, see). Though it has become evident that high-level factors influence low-level features; that is, a top-down process substantially influences gaze guidance (Schomaker, Walper, Wittmann, & Einhäuser, 2017). State-of-the-art computational models of saliency use low and high-level features of an image as input and output a predicted topographical map of how salient each area of that image is to a human viewer (Koehler et al., 2014; J. Xu & Yue, 2014).

Turning to saliency in videos, little research to date has been carried out on molding dynamic saliency in videos (e.g., Sidaty, Larabi, & Saadane, 2017), mainly because of the complication arising from motion and temporal information in video streams (W. Wang, Shen, & Shao, 2017). Saliency in video data is dynamic and strongly related to changes in both spatial and temporal features. For example, in dynamic scenes, a simple moving object can draw the observer's attention and the moving object does not have to be different from its surroundings by means of the *Gestalt Principles* (Kavak, Erdem, & Erdem, 2017).

In the last few years, there has been some progress in modeling saliency in videos, for example, through the integration of low-level (e.g., static saliency, faces, and texts) visual features with more dynamic ones (e.g., motion) (Itti & Baldi, 2005; Le Meur, Le Callet, & Barba, 2007; Mahapatra, Gilani, & Saini, 2014; Nguyen et al., 2013). Many dynamic models adopt the same approach; extracting spatial and temporal saliency maps and then fuse them to obtain a final spatiotemporal saliency map (see, Kavak et al., 2017, for a recent review). Recently, there has been an improvement in dynamic saliency prediction using deep learning approaches (N. Liu & Han, 2018). These systems model dynamic saliency using a wide range of features that represent saliency as it enfolds in space and time.

#### 3.4.3.4 Human Face and Gaze

Human faces are attention magnets. Indeed, both anecdotal and scientific evidence, suggest that human faces attract more attention than any other object in the visual scene. Research has shown that when presented with other objects, our attention is automatically drawn towards human faces before it shifts to other objects. More interestingly, we also pay more attention to human-like faces, such as faces of humanoid and animated pedagogical agents (Louwerse, Graesser, McNamara, & Lu, 2009). There is ample evidence that the human face is rapidly and easily detected when presented among other objects

(Hershler, Golan, Bentin, & Hochstein, 2010). However, the question of how and under what conditions faces guide attention is constantly debated in face perception literature.

In the field of multimedia learning, it is generally assumed that showing the instructor face (and bodily gestures) in video affords useful nonverbal cues, e.g., mutual gaze, facial expression, and gestures, to learners which may aid them in processing and resolving disambiguation in the verbal information that is narrated by the instructor. As verbal and visual information is processed in two separate channels, as suggested by the *Dual Coding Theory* (J. M. Clark & Paivio, 1991), integrating both cues may result in an enhanced comprehension of the material. Another advantage of showing instructor's face is suggested by the *Social Presence Theory* which postulates that different media (e.g., face-to-face, telephone, text, and video) convey differing degrees of social cues that make learners to feel more or less psychologically present with one another. From this perspective, the presence of the instructor's face in video replicates the social aspects of human interaction and makes the learner feel as if they are interacting with another person (J. Wang & Antonenko, 2017).

In a recent review of the image principle of the *Cognitive Theory of Multimedia Learning*, Mayer (2014) asserted that adding the speaker's face in instructional materials does not necessarily lead to strong, positive effects on learning. But, as noted by J. Wang and Antonenko (2017), the image principle was tested using low-embodied or high-embodied animated pedagogical agents, and therefore its effect may not be extended to actual human faces.

Recently, some studies have explored if showing an instructor's face onscreen leads to better learning gain, information retention, and recall (e.g., Kizilcec, Papadopoulos, & Sritanyaratana, 2014; Stull, Fiorella, & Mayer, 2018; van Wermeskerken & van Gog, 2017; J. Wang & Antonenko, 2017). The findings of this line of work have generally suggested that while the presence of instructor's face is preferable for learners and often perceived to be effective, there is no significant gain on learning due to the presence of the instructor face per se. For example, Kizilcec et al. (2014) investigated the impact of learning from video lectures with and without instructor face present. Eye-tracking data indicated that learners spent about 41% of their time looking at the face and switched between the face and slide every 3.7 seconds, but they did not perform significantly better on short-term or mid-term recall tests compared to learners in the control condition.

J. Wang and Antonenko (2017) examined how the instructor presence or absence would influence learning, visual attention distribution, and perceived learning, mental effort,

and satisfaction when viewing mathematics instructional videos at easy and difficult levels of content complexity. The results indicated that displaying the instructor image on the video attracted considerable attention, particularly when learners viewed the video on an easy topic. The researchers ascribed this finding to the relatively low intrinsic cognitive load imposed by videos on an easy mathematics topic which allowed learners to attend more to the instructor. Concerning the impact on learning from the videos, the study reported no significant difference in transfer learning—the ability to apply what has been learned from the video to a new context—for the easy and difficult topic, though the study participants' recall of information from the video was better for the easy topic when instructor's face was present.

Taken together, the findings of previously discussed studies seem to suggest that although the presence of instructor's face in instructional videos may provide useful nonverbal and social cues (e.g., facial expressions, eye gaze) to the learners, paying attention to this information could, nevertheless, impose unwanted extraneous cognitive load on the learners and distract them from attending to and learning from the video content, especially when the topic of the video in itself is already cognitively demanding (J. Wang & Antonenko, 2017).

While the previously discussed studies have primarily focused on the entire face, a recent strand of research has begun to explore how instructor's eye movements could guide learner attention when learning from video demonstrations (Ouwehand, van Gog, & Paas, 2015; Van Gog, Verveer, & Verveer, 2014; van Wermeskerken, Ravensbergen, & van Gog, 2017). Demonstration videos (or how-to-videos) differ from other types of instructional videos (e.g., recorded lecture) in that they show an instructor demonstrating how to perform a certain task. In this type of instructional video, it is reasonable to assume that closely observing the instructor's face is not as important as paying attention to the task he/she is demonstrating.

In a study reported by Van Gog et al. (2014), 43 learners watched a video on how to solve a puzzle twice and attempted to solve the problem on their own after each viewing of the video example. The participants were divided into two groups; the first group watched a video example in which the instructor's face was visible, and the second group did not see the face (only the hands solving the puzzle problem). The results showed that the instructor's face attracted considerable attention (23% of all fixations on the first viewing and 17% on the second viewing) and learners who watched a video example with the instructor's face visible performed better than the learners in the invisible instructor face condition. The researchers suggested that the positive effect on learning is probably due

to learners automatically following the instructor's gaze as she was performing the task. In a recent study, van Wermeskerken et al. (2017) attempted to replicate the finding of Van Gog et al. (2014), using a different demonstration task (learning to build a molecule). van Wermeskerken et al. (2017) replicated the two conditions in but also a third condition in which the instructor's face is visible but offer no gaze guidance (the model looking straight into the camera and not at the molecule). Analysis of the data suggested that learning was neither facilitated nor compromised when the instructor's face is shown. Further, the eye movement data demonstrated that participants can efficiently distribute their attention between the instructor's face and the task he or she is demonstrating.

Looking at someone's face while talking can enhance speech comprehension for both L1 and L2 listeners, especially in challenging and adverse listening conditions (e.g., Navarra & Soto-Faraco, 2007). Research has shown that visual speech—which typically consists of movements of the tongue, teeth, and lip movements—provides L2 listeners with phonological information of the speech signal which enhance comprehension (Navarra & Soto-Faraco, 2007)

#### 3.4.3.5 Image Complexity and Visual Clutter

Several computational measures of image complexity have been proposed based on, for example, information theory (Rigau, Feixas, & Sbert, 2005) and fuzzy approaches (Cardaci, Di Gesù, Petrou, & Tabacchi, 2009). Perhaps the simplest and widely used measure of visual complexity relies on the resulting image size when applying compression algorithms, like JPEG, on the original image. The rationale behind this approach is that the more complex the image is, the more file size it needs in computer memory. Other computational measures estimate visual complexity by computing low-level features, e.g., color, edges of objects, intensity variation, and so on. Although these measures were originally developed for computer vision and image processing tasks, cognitive researchers found some of these measures to correlate with human-based subjective evaluations of visual complexity (Machado et al., 2015; Yu & Winkler, 2013).

Recently, it has been postulated that image complexity is a multidimensional construct; that it is not only determined by the number of elements or structural differences in the image but also by the perception of the observer (Gartus & Leder, 2017). To develop a multi-dimensional model that could take combine several features, some researchers have employed machine learning techniques to predict the human perception of visual complexity (Gartus & Leder, 2017; Machado et al., 2015). In their study, Machado et al. (2015)

compared the performance of several computerized metrics—based on image compression error and Zip's law—and machine learning-based systems techniques. The findings demonstrated that incorporating multiple image features in a machine learning-based system yielded the most accurate predictions of human perceived visual complexity. Of several image metrics investigated, the researchers found that edge density and compression error metrics are the best predictors of the participants' ratings of visual complexity.

Closely related to image complexity is the notion of visual clutter (Corchs et al., 2016). According to Rosenholtz, Li, and Nakano (2007), clutter "is the state in which excess items, or their representation or organization, lead to a degradation of performance at some task" (p. 3). Intuitively, the more cluttered an image is, the more information it contains. Therefore, visually cluttered images are more likely to impede visual object processing and result in performance and attentional costs. Empirical studies have shown that cluttered images or displays increase memory load, delay visual search, negatively affect object recognition, reading, and linguistic processing (Coco & Keller, 2012). In the literature, there are several approaches and algorithmic methods to measure clutter in visual displays (see Moacdieh & Sarter, 2015, for comprehensive review). Two common algorithms are the feature contestation and subband entropy algorithms proposed by Rosenholtz et al. (2007) and Rosenholtz, Li, Mansfield, and Jin (2005).

#### 3.4.3.6 Gestures

Gestures are spontaneous movements of the hands or arms that are tightly synchronized with speech (McNeill, 1992). With decades of research on gestures, it has become increasingly clear that gestures can enhance comprehension of spoken language, though how and under what conditions is still a matter of debate (Dargue & Sweller, 2020a). Several studies have shown that gestures benefit both language speakers and listeners (e.g., Dargue & Sweller, 2020b; Dargue, Sweller, & Jones, 2019; Hostetter, 2011; Sueyoshi & Hardison, 2005). For speakers, gestures help organize thinking, facilitate retrieval of words from memory, and reduce cognitive burden. For listeners, they can enhance comprehension of a spoken message (e.g., Cassell, McNeill, & McCullough, 1999). In this way, gestures fulfill two primary functions; extending our ability to express complex ideas and helping us understand other people's speech (McNeill, 1992). In a meta-analysis study, Hostetter (2011) reported that gestures do provide a significant, moderate benefit to listener's comprehension when compared to observing no gestures at all. The benefit of gestures on speech comprehension has been recently confirmed by a meta-analysis study

TABLE 3.6: Visual Complexity Features, their Impact & Supporting Empirical Evidence

| | **Hypothesized Impact** | **Empirical Evidence** |
|---|---|---|
| Visual clutter | The more cluttered an image is, the more difficult it becomes to locate and identify a target. | (Coco & Keller, 2012; Donderi & McFadden, 2005; Neider & Zelinsky, 2011; Wolfe, 1994b) |
| Visual saliency | Given that human attention is limited, only silent objects are processed while other objects are ignored | (Clarke, Coco, & Keller, 2013) |
| Object color | Color guide visual attention and may help in object recognition, especially when objects are color diagnostic (e.g., strawberry is strongly associated with the color red). | (Bramão, Reis, Petersson, & Faísca, 2011; Tanaka, Weiskopf, & Williams, 2001) |
| Gestures | Gestures generally facilitate comprehension, especially in degraded or noisy listening conditions. However, the realization of gestures may also vary from one culture to another which makes interpreting their meanings less straightforward. | (Drijvers, Vaitonytė, & Özyürek, 2019; Sueyoshi & Hardison, 2005) |
| Human Face | Human face automatically attracts more attention than any other objects in the visual scene. When there is no task being demonstrated, showing a human face can be helpful, but, if a task is being demonstrated, showing the model's face may lead to a division of attention. | (Hershler et al., 2010; Ouwehand et al., 2015) |
| Motion | Motions, especially abrupt and sudden motions, rivet the viewer's attention involuntarily, hence they might be distracting if they are not semantically relevant to the current task. | (Mital, Smith, Hill, & Henderson, 2011) |

(Dargue et al., 2019). It should be noted that gestures are of different types (McNeill, 1992) and they vary in the level of support they provide to listeners.

Interestingly, gestures help speech comprehension, even though gestures and speech represent two expressive modalities, and producing or processing them simultaneously should in theory increase cognitive load. One plausible explanation for this is that speech and gestures are two sides of the same coin, and we rely on both for interpretation (Kelly, Özyürek, & Maris, 2010). In a recent study examining the neural bases of speech and gesture interactions, it has been found that the semantic processing of gestures and language rely on overlapping resources in the brain (Holle, Obleser, Rueschemeyer, & Gunter, 2010; Kelly et al., 2010; Özyürek, Willems, Kita, & Hagoort, 2007).

Support for the role of gestures in facilitating speech comprehension, both in clear and degraded speech, are well-documented in the literature. When combined with speech, for example, gestures aid in disambiguating the meaning of difficult speech (Hostetter, 2011; Sumby & Pollack, 1954). Several studies also demonstrated that observing iconic gestures that are redundant with speech enhances comprehension compared with observing no gestures at all (Dargue & Sweller, 2018; Dargue et al., 2019).

Research in gestures and their communicative functions points to the fact that gesture is an intricate communicative system in its own right, with several types of gestures fulfilling distinctive communicative functions (Cassell et al., 1999). From a semiotic perspective, gestures "constitute a rich expressive resource available to speakers who deploy them with speech in sophisticated, systematic and non-trivial ways" (Gullberg, 2010, p. 76). While the majority of gesture types co-occur with speech, emblematic gestures do not accompany speech and are expressive in themselves (Ekman & Friesen, 1969). The realization of gestures is linguistically and culturally bond to the speech community in which gestures are used; hence, it stands to reason that gestures may also be an obstacle for L2 listeners.

The role of gestures in L2 listening has received limited attention in second language literature and often played a secondary role in the study of verbal language in spoken interaction (Dahl & Ludvigsen, 2014). However, there seems to be an increasing interest in exploring the role of gesture in L2 language learning in the last few years. In a study conducted by Sueyoshi and Hardison (2005) to instigate the contribution of gestures and facial expression to spoken comprehension, low-intermediate and advanced learners of English benefited from watching video recorded lecture in which the presenter's face and his face and gestures are shown but not when they only heard the audio recording of

the lecture. Similar findings are also reported in Dahl and Ludvigsen (2014) who found substantial benefits for gestured speech for L2 listeners.

Research also suggests a relationship between the learner's language proficiency and age, and how likely he or she would benefit from gestured speech. For instance, Mohan and Helmer (1988) found that when compared to native children, L2 children did not benefit from gestures speech. Mohan and Helmer (1988) attributed this finding to the children's low language proficiency. In a recent study comparing how L1 and advance L2 listeners allocate their attention to visual speech and gestures, Drijvers et al. (2019) found that L2 language proficiency impacts overt visual attention to articulators, which in turn result in different visual benefits for native versus non-native listeners.

#### 3.4.3.7 Cinematographic Effects

Perhaps the most compelling evidence for how human visual attention work does not come from laboratory and clinical experiments, but rather from film studios. To tell an engaging story, filmmakers have for long been experimenting with different cinematographic techniques to create, normally from hundreds or thousands of shots, a perceptually coherent sequence of events (Bordwell & Thompson, 2001; Zacks, 2003). Through closely observing audience reactions to different editing techniques, cinema pioneers were able to develop an informal theory of human visual cognition and perception which allowed them to creatively disassemble and reassemble different views of a scene without disturbing the development of a film's narrative (Levin & Baker, 2017). This implicit theory of audiovisual narrative comprehension has, of course, evolved over the years and translated into more sophisticated editing techniques and principles.

One set of powerful filmic techniques is known as *continuity editing rule* or *Hollywood style*. Continuity editing includes techniques such as cuts, dissolves, fades, or wipes which now permeates a wide range of dynamic visual media. But by far the most used editing technique is straight cuts, counting for approximately 99 percent of transitions in contemporary films (Cutting, DeLong, & Brunick, 2011). In films, the primary purpose of cuts is to elide shots, typically token in different places and time, in a way that a sense of continuity is maintained (Bordwell & Thompson, 2001). In professionally edited films, viewers' attention is guided or directed towards the most informative portions of the screen during each successive shot to encourage comprehension (Kirkorian & Anderson, 2018). In fact, through tracking viewers' eye-movements, researchers observed that a strong tendency among film viewers to look at the same place on the screen at the same time as each other

(Dorr, Martinetz, Gegenfurtner, & Barth, 2010; Mital et al., 2011), a phenomenon called as *"attentional synchrony"* (Kirkorian & Anderson, 2018). It is generally assumed that when viewers' attention is synchronized, they in the main arrive at the same understanding. But in some other cases, however, filmmakers may also want to break or discontinue the flow of the narrative. One way to achieve this is through the use of precipitous cuts, such as fade-outs and wipes, to signify the transition from one episode to another in the narrative.

While invisible cuts help narrative runs smoothly in films, in educational videos, cuts may need to be abrupt to signify the introduction of new information to the viewers, i.e., a slide-based lecture. With every change in shots, it is assumed that learners need to deploy more cognitive resources to process this new information and integrate it with the old one in their memory (Magliano & Zacks, 2011). A corollary of this is that in instructional videos precipitous cuts should be used judiciously and only be used when there is a change in content. Also advisable is segmenting videos into manageable chunks, especially if segments make separable different steps or procedures in demonstrational videos(Spanjers, van Gog, & van Merriënboer, 2010). This is because video segmentation reduces the lability of the information and allows learners to absorb what they have already watched.

It is quite interesting how these superficial techniques work and how they deceive the human perception of narratives. Films and other types of visual narratives have intrigued many cognitive scientists, and there is a growing body of research exploring the cognitive science of cinema (e.g., Levin & Baker, 2017; Levin & Simons, 2000). One of the most important questions organizing visual narrative research is to which the degree of the understanding of visual narrative depends on medium-specific skills versus everyday perceptual and cognitive skills. There have been two alternative responses to this question. The first one posits that films are merely a reflection of reality and therefore do not require special cognitive processes(e.g., J. Anderson, 1996; Levin & Simons, 2000). As such, film viewers normally have no difficulty understanding films despite their inherent complexities (Zacks, 2003). The second view posits that films deviate from real-world perceptual experience (Mnsterberg, 1970). For example, through the use of cinematographic techniques, e.g., flashback, films are liable to depict different realities. These experiences have no counterpart in everyday life. A middle view seems to be right; film viewers rely on both common or natural perceptual processes and learned medium-specific skills to understand moving pictures. In a series of studies, Schwa and associates found evidence

that suggests first-time film viewers struggle to make sense of the narrated story, suggesting that film viewers also need to be familiar with filmic conventions in order to fully comprehend films (Ildirar & Schwan, 2015; Schwan & Ildirar, 2010).

#### 3.4.3.8 Events and Actions

When watching dynamic moving pictures, viewers perceive the continuous stream of audiovisual information as a series of discernible units or events" (Zacks & Tversky, 2001). This process is called *event structure perception* and the units or events are normally defined as "a segment of time at a given location that is conceived by an observer to have a beginning and an end." (Zacks & Tversky, 2001). Illustrative examples of an event include making a cup of coffee, driving a car, opening a door, or folding the laundry. Event cognition researchers have studied how people segment activity as it is happening and what are the underlying cognitive mechanisms behind this process. In a classical study, Newtson and Engquist (1976) developed a unit marking procedure, also called an event segmentation task, in which participants are presented with a video clip of an ordinary task, like preparing breakfast, and asked to press a button when they believe one event ends, and another one begins. The findings revealed that participants were remarkably consistent in their responses and reliably identified points in the clips which indicate event boundaries. The same finding has been found by other studies that explore event segmentation in videos of everyday activities (Boggia & Ristic, 2015), visual and written narratives (Bailey, Kurby, Sargent, & Zacks, 2017), and dance movements (Bläsing, 2014).

Recent research on event cognition has revealed important aspects of how humans parse a continuous stream of behavior into meaningful and recognizable events, and how this affects memory and cognition (see, Radvansky & Zacks, 2017, for a review). A theoretical account of how everyday events are created, structured, and remembered is described in the *Event Horizon Model* (Radvansky & Zacks, 2014). The model consists of five principles, the first of which is relevant here. The principle states that humans automatically parse into meaningful units. This principle is further explained in the *Event Segmentation Theory* (EST: Zacks et al., 2007). According to the EST, the effective segmentation of observed events is fundamental for understanding and memorizing these events.

There is considerable empirical evidence supporting the role of event segmentation in

TABLE 3.8: Cinematographic Complexity Features, their Hypothesized impact and Supporting Empirical Evidence

| | **Hypothesized Impact** | **Empirical Evidence** |
|---|---|---|
| Camera changes & cuts | A camera change signifies a change in information to the viewer, hence it elicits an automatic allocation of additional cognitive resources. | (Hendriks Vettehen & Kleemans, 2015; Lang & Geiger, 1996; Reeves et al., 1985) |
| Production style | Different instructional video production may have a different impact on learning and some learning concepts seem to be learned better in one video style but not in another (e.g., teaching math in Khan academy style, and coding in screencasted video tutorial). | (Mayer, Fiorella, & Stull, 2020) |

video understanding (Biard, Cojean, & Jamet, 2018; Boggia & Ristic, 2015; Magliano, Radvansky, Forsythe, & Copeland, 2014). Researchers have also explored how video edits—that reinforce the segmentation of events—can improve memory (Schwan, Garsoffky, & Hesse, 2000). Gold, Zacks, and Flores (2017) manipulated visual and auditory cues in a set of movies to ascertain whether segmentation enhances subsegment memory of events. For each movie clip, cues were placed either at event boundaries or event middles, or the clip was left unedited. The results showed that segmentation, especially at the event boundaries, can improve subsequent memory of everyday events. In a recent study, event segmentation was found to contribute to memory formation and recall of information up to one month later and that individual differences in event segmentation are related to differences in memory over long delays (Flores, Bailey, Eisenberg, & Zacks, 2017).

#### 3.4.3.9 Video Production Styles

Although some fundamental video production techniques, e.g., cuts, are common across different video genres, some editing techniques are unique to certain video genres. Collectively, these techniques are what define video genres and make them distinguishable from each other. Instructional videos, for example, differ drastically from movies or documentaries. To maximize learning and minimize extraneous load, instructional videos

are slow-paced, have fewer shots and camera movements, and do not involve extensive use of graphics or music. Within the broad genre of instructional videos, we can also talk about sub-genres or styles which vary in their learning design choices. Attempts to categorize instructional videos are seen in recent proposed topographies of instructional video styles (e.g., Crook & Schofield, 2017; Winslett, 2014). Among popular instructional video, styles are talking head, picture-in-picture, and khan academy style. Given their differences, these video styles would perhaps vary in their learning potentials.

When making a video for an instructional purpose, video content developers and instructional designers need to make several design decisions, for example, whether to show the instructor's face or hide it, to add texts on screens or not and so forth. These decisions can dramatically impact learning from videos. According to Crook and Schofield (2017), "video is a medium whose properties and impact depend on the creative shaping of designers, editors, and producers" (p.57). Although instructional video developers and creators ordinarily strive for making their video contents more effective and conducive to learning, Fiorella and Mayer (2018) noted that instructional videos, by and large, are designed based on intuitions, not on learning theories and documented scientific design principles.

In learning design science and multimedia learning literature, researchers have been interested in ascertaining how people learn from dynamic visualizations and videotexts and how this learning can be boosted (Fiorella & Mayer, 2018). The outcome of such research effort was a set of recommendations or designing principles of what seems to work and what not (e.g., Mayer et al., 2020). For instance, the *perspective principle* suggests that people learn better from demonstration video when it is filmed from a first-person perspective rather than a third-person perspective. In first-person perspective videos, the camera is placed on or above the instructor's shoulder or forehead as she or he demonstrates an activity, whereas, in the third-person perspective video, the camera is placed across the demonstrator. Another design principle is the dynamic drawing principle which suggests that people learn better from a video lecture that shows the instructor drawing graphics while lecturing rather than referring to already drawn graphics. Evidence for this principle comes from recent studies (Fiorella, Stull, Kuhlmann, & Mayer, 2019; Fiorella, Van Gog, Hoogerheide, & Mayer, 2017) in which learners are found to benefit from observing instructor's dynamic drawings.

Given the wide range of video instructional styles and their use of different design principles, it is plausible to assume that, all things being equal, production style may have different affordances and impact on learning and that some learning concepts are better

learned in one production style but not in another. Further, learners' preference for and familiarity with a particular video style may impact their learning. However, it remains unclear how and what video production styles are more effective for language learning.

### 3.4.4 Summary of Visual Complexity

In this section, I reviewed research on visual complexity and looked at some attributes of video imagery that seem to intervene with and hamper videotext understanding. Admittedly, the review is, by no means, comprehensive, but it nevertheless highlights key aspects of visual complexity in videotexts. Central to the notion of visual complexity discussed here are the selectively and limited capacity of the human attention system. For example, aspects related to object saliency, and abrupt change in color and motion impact videotext processing. To put it differently, the transient nature of videotext makes it difficult for language learners to process and coordinate information coming from both visual and verbal channels simultaneously.

## 3.5 Chapter Summary

In this chapter, I reviewed L2 video studies in the past two decades to ascertain how L2 researchers assessed or controlled for the difficulty in their video research. The findings indicated a clear gap in the literature on our understanding of videotext complexity and what possibly makes videotext difficult for language learners. There was also a palpable need for automated videotext difficulty assessment tools for use in language teaching, learning, research, and assessment, akin to those already available for spoken and written texts.

To develop a conceptual understanding of videotext difficulty and complexity, I first reviewed key concepts and principles from the field of text/discourse comprehension and discussed some recent development in visual narrative comprehension research. Then, I defined videotext difficulty and circumscribed its interaction with seemingly similar constructs, videotext complexity, and comprehensibility. Then, in the final section of the chapter, I reviewed and synthesized extant empirical evidence for the role of acoustic, phonological, lexical, syntactic, discoursal, and visual complexity on videotext understanding. While the prolificity of verbal and visual complexity predictors may be helpful to understand videotext difficulty, their sheer number and inextricable interactions

nonetheless present serious practical challenges. In the next chapter, I will discuss how the use of machine learning can help in gaining useful insights from complex and unstructured data as videotexts.

# Chapter 4

# Machine Learning for Videotext Difficulty Assessment

In the previous chapter, I reviewed a substantial body of literature on verbal and visual complexity from diverse fields of research, with the aim of gathering and synthesizing empirical evidence on potential predictors of videotext difficulty. While these complexity features may help us understand the multifarious facets of videotext complexity and how they might contribute to videotext difficulty, their sheer number, nonetheless, poses serious practical challenges (e.g. model's overfitting) for developing automated measures of videotext difficulty assessment.

In this thesis, I adopted methods and algorithms from the field of machine learning and AI to develop more reliable and robust predictive models of videotext difficulty. The chapter provides a brief introduction to key concepts and notions, particularly within the context of supervised machine learning; whereas clarification on methods employed in each study is detailed in the relevant chapters. Later, in this chapter, I outline the overall research design and describe the objectives of the four studies conducted as a part of this thesis.

## 4.1 Introduction

This thesis is set to explore a systematic way to computationally explain and predict videotext difficulty for use in second language research, assessment, and learning. Computational approaches to text difficulty assessment have progressively evolved over the

last two decades, and contemporary systems leverage the recent advancements in machine learning as well as the increasing availability of large-scale datasets for developing more reliable measures of text difficulty (Collins-Thompson, 2014). In the following sections, I provide a brief description of key concepts and methods commonly used in the development and validation of machine learning-based systems for automatic text difficulty assessment. I will also point out the usefulness of predictive modeling in understanding videotext difficulty.

## 4.2 Machine Learning: Basic Notions

The field of machine learning is a sub-field of artificial intelligence and it has gained a currency in the last two decades due to the increase in computing power and the need for devising new algorithms for harnessing and gleaning new insights from big data. The widespread use of machine learning methods in scientific research is in part due to the development of easy-to-use machine learning libraries and tools, along with the availability of down-to-earth introductory books and online video tutorials. Two such excellent books in statistical and machine learning are *An Introduction to Statistical Learning* by G. James, Witten, Hastie, and Tibshirani (2013) and *The Elements of Statistical Learning Data Mining, Inference, and Prediction* by Hastie, Tibshirani, and Friedman (2009), from which I draw generously in the following sections.

As I alluded to in Chapter 2, researchers in the field of text difficulty assessment have primarily used supervised machine learning algorithms for the task of assessing difficulty in written and spoken texts. Similarly, in this thesis, our task of assessing and predicting videotext difficulty falls in the supervised learning domain. In order to motivate our discussion of supervised machine learning, I will use the task of predicting text difficulty, as an example. In this task, let's suppose that, $X_1, X_2 \ldots X_p$ are some characteristics of the text, and $Y$ is a variable that reflects the difficulty level of the text. When the output or dependent variable, $Y$, is continuous number, our task is a *regression* problem; whereas if our dependent variable is discrete or categorical, our problem becomes a *classification* problem. Further, let's assume that there is a relationship between $\vec{X}$ and $Y$ and our goal is to learn a mathematical function, $f$, that maps our input variables to the output variable in a way that can help us understand their relationship. Formally, the function $f$ can be written in the very general form as in Equation:

$$Y = f(X) + \epsilon \tag{4.1}$$

In this equation, $f$ is some fixed but unknown function of $X$ and $\epsilon$ is a *random error* term. The error term should be independent of $X$ and has a mean of zero. In machine learning, there are many approaches to estimate the unknown function $f$ based on the observed data points (or instances), which can be generally characterized as either parametric or non-parametric (G. James et al., 2013). Parametric methods make a prior assumption about the shape of $f$. A very common assumption is that $f$ takes a linear form, written formally as:

$$F(X) = \beta_0 + \beta_1 X_1 \tag{4.2}$$

The term $\beta_0$ is an intercept, also known as bias in the field of machine learning.

Since we assumed the $f$ is a linear function, then the task of finding the unknown parameters becomes much more constrained. This is because we only need to estimate a handful set of parameters $X_1, X_2 \ldots X_p$. To estimate those parameters, we need a sample of data to fit the model. This data is often called a *training data* because we want to use it for training or teaching our algorithm how to estimate $f$. In the case of a linear model, there are several approaches to fit the model, but by far the most popular approach is ordinary least square (OLS) and it helps to reduce the gap between the predicted, $\hat{Y}$, and actual values (or ground truth), $Y$. The linearity assumption in parametric approaches is both their strength and weakness. Assuming a linear form for $f$ simplifies the problems of estimating $f$ because it is generally much easier to estimate a set of parameters in the linear model than it is to search for an estimate of $f$ in an entirely arbitrary function space. However, one key drawback of linear approaches is that the functional form used to generating $\hat{f}$ is more likely to be too far from the true $f$, in which case the fitted model will not fit the data very well. In other words, there is a trade-off between a model's prediction accuracy and its interpretability. I will return to this important issue later. For now, let's discuss non-parametric approaches.

Unlike parametric or linear approaches, non-parametric approaches do not make any explicit assumption about the shape or functional form of $f$, but instead, seek to find the best estimate of $f$ from a wide range of possible potential functional forms. In this sense, non-parametric approaches are more flexible and can lead to a more accurate estimate of $f$. But to arrive at the most accurate function $f$ that fits the data very well, a large

number of observations is required, far more than what is typically needed for parametric function. One might ask if the problem of estimating $f$ in non-parametric approaches is more challenging why not use a parametric approach instead. The choice between the two approaches chiefly depends on why we are estimating $f$ in the first place. Generally, there are two main objectives for estimating $f$; making inference or prediction. Determining the primary objective of using statistical modeling before embarking on any research project is critical, as it helps researchers in issues related to choosing the most appropriate machine learning algorithm for the task, evaluation criteria, determining statistical power, and whether collecting a large dataset is necessary or not (Shmueli, 2010).

### 4.2.1 Explanation and Prediction

The main objective of statistical modeling has always been a controversial topic in the philosophy of science literature (e.g., Hempel & Oppenheim, 1948). In essence, there are primary objectives of statistical modeling in scientific research: to *explain* a phenomenon or to *predict* its occurrence in the future. Researchers across different scientific fields tend to focus on achieving one of the two objectives. In social sciences, researchers are predominantly interested in *explaining* a particular phenomenon of interest so they rely on inferential statistics for testing causal hypothesis within a theoretical account of the phenomenon, and rarely, if ever, they apply predictive modeling techniques to their data. In applied and predictive sciences, e.g., natural language processing and bioinformatics, the reliance on casual theory is much weaker and researchers in those fields preferentially select predictive modeling over explanatory modeling because the type of problems and data they work with do not lend itself to causal explanation (Shmueli, 2010).

It is commonly assumed that the two approaches or paradigms reflect two competing goals of doing science: causal explanation and empirical prediction (Shmueli, 2010). This assumption often led to debates about, or even prejudice against, what approach is deemed scientific and what is not. Nagel (1961) asserted that "the distinctive aim of the scientific enterprise is to provide systematic and responsibly supported explanations" (p. 15). Douglas (2009) argued that "explanation and prediction are best understood in light of each other and thus … should not be viewed as competing goals but rather as two goals wherein the achievement of one should facilitate the achievement of the other" (p. 445). Under this assumption, if an explanatory model sufficiently describes the questioned (causal) relationship in a dataset, including all the moderating and mediating variables that govern when and to what extent they each influence behavior, the model, at least in

principle, should be capable to predict with reasonable accuracy its occurrence in future data (Yarkoni & Westfall, 2017).

However, although explaining and predicting a behavior may be philosophically compatible, statistically, it is not always true that the model that best approximates the data generating processes that produce an observed behavior is the one that best predicts future behavior (Shmueli, 2010; Yarkoni & Westfall, 2017). In many cases, a theoretically plausible and elegant model fails miserably to predict a phenomenon whereas a very complex model that is incomprehensible to humans succeeds in accurately predicting outcomes of interest (Yarkoni & Westfall, 2017). Thereby, scientists often make a deliberate choice as to whether to provide a more plausible theoretical explanation or more accurate empirical prediction of outcomes of interest. This in turn affects the entire process of statistical modeling i.e., the choice of algorithms, sample size, and the analysis and reporting of the results (Shmueli, 2010).

Machine learning algorithms vary on their flexibility and hence their interpretability. As we discussed earlier, non-parametric methods are generally better in capturing the patterns in the data because of their high flexibility, but they lack interpretability. When the goal is to explain the relationship between the variables, using a simple and less flexible machine learning algorithm (linear approaches) is preferable. For example, OLS regression is relatively inflexible but it is quite interpretable. When the goal is to predict new feature values, then a more flexible method is preferable. For example, polymodal regression and generalized additive models (GAMs) are capable of modeling nonlinear (e.g., curve) relationships between predictor and response variables (G. James et al., 2013). Highly flexible methods, e.g., bagging, boosting, and artificial neural networks can lead to more accurate prediction, especially when trained on a massive dataset. However, highly flexible algorithms do not always lead to accurate predictions. It is very plausible that a more restrictive model performs better than a less restrictive or complex algorithm. This is due to several reasons, but by far the most common one is *overfitting*; that is, the model learned the data very well. Because the model fits too closely to the training dataset, it becomes less generalizable to other observations. Overfitting can also happen when the number of features exceeds the number of observations in the dataset. To mitigate the impact of overfitting, regularization methods, such as LASSO or L1 and Ridge or L2 penalty terms, may be used. Contrary to overfitting, a model may be *underfitting* the data, that is it does not capture meaningful patterns in the data. As it might be already suggested, both overfitting and underfitting render unreliable machine learning models and researchers should make sure their models strike a good balance.

When the accuracy of the prediction is the goal, then the model's interpretability is not a concern. In other words, it does not matter if the predictors in the model have a logical relationship with the dependent variable. Nor should we be concerned about what algorithms, simple or complex, we ended up choosing for the task. In this case, the use of "black box" models is perfectly justifiable. One approach commonly used by statistical and machine learning researchers to optimize prediction accuracy at the expense of interpretability is ensemble learning. In this learning paradigm, multiple machine learning models (commonly called "weak learners" or "base models") are combined to get better predictive accuracy (Witten, Frank, Hall, & Pal, 2016). The main hypothesis behind this approach is that when several "weak" learners or models are correctly combined to solve the same problem they often lead to better prediction accuracy than a single model (Z. Zhou, 2012).

To sum, there is a tradeoff at play between prediction accuracy and model interpretability. That is, more complex models often yield better estimates of $f$ but render 'black box' models which are difficult to interpret. Depending on the primary objective of model building (explanation or prediction), researchers should select machine learning algorithms that are the most appropriate for the task. In this thesis, we applied both *explanatory* and *predictive* modeling approaches for making *inferences* and *prediction* of video difficulty, respectively. For understanding how verbal complexity (Chapter 6) and visual complexity (Chapter 7) relate to video difficulty, I use more restrictive algorithms (e.g., OLS regression and LASSO). In Chapter 8, I rely on more flexible, e.g., bagging, boosting, support vector machines with non-linear kernels, for predicting video difficulty using multimodal complexity features.

### Hyperparameters Tuning

Some advanced machine learning algorithms require a set of values or hyperparameters to be optimized or tuned for optimal performance, e.g., the number of trees in a random forest algorithm. The best hyperparameter space for a machine learning algorithm can be found using different searching methods such as grid searching, randomized grid search, or Bayesian-based optimization techniques. A popular approach for finding the optimal values for these hyperparameters is via a *grid search* method. A grid search is a hyperparameter optimization technique that searches exhaustively through a manually specified subset of the hyperparameter space of the targeted algorithm.

### 4.2.2 Evaluating a Learned Function

Regardless of what machine learning approach, parametric or non-parametric, was chosen for the task, we need a reliable method to evaluate the accuracy of the learned model. In most cases, we would need to fit several models to the dataset—because there is no single model that works best in all dataset and tasks (this is commonly referred to as *no free lunch theorem* in the field of machine learning)—and compare their performance against each other. Specifically, in a regression setting, for each input pairs $(X, Y)$, we need to quantify how close the model prediction, $y$, is from the actual or truth value, y, for that observation. There are different evaluation metrics for regression problems, but the most common one is mean square error ($MSE$), given by the following equation:

$$MSE = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.3}$$

The average value of MSE will be small if the model predictions are close to the true values and by contrast, it will be large if there is a substantial difference between the model predictions from the true values. and root mean square error (RMSE). One distinct advantage of RMSE over MSE is that it expresses the model prediction error in units of the variable of interest. The RMSE is a

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n}} \tag{4.4}$$

One useful property of RMSE is that its values are of the same dimension as the predicted values, hence it is more interpretable than other metrics (Witten et al., 2016).

If our goal is to train a model for future prediction, a higher model's accuracy in the training set is not an indicator of a good model (Witten et al., 2016). That is, we are not confident if the trained model would also perform well on unseen observations, which is the primary objective of developing a predictive model in the first place. The question then is how we can ensure that our trained models can yield similar or consistent results when applied to new observations. A common validation approach is to divide the dataset into a training and testing set. As their names imply, the training set is used to train the model (estimating its unknown parameters) and the test set is held out for evaluating the model accuracy. Typically, two-third of the dataset is used for training the machine

learning model and the remaining one-third is preserved for testing the accuracy (Witten et al., 2016).

This is because the training set may not be representative of the problem the model tries to learn. One way to alleviate this problem is through shuffling the dataset before splitting the dataset into training and test set. One caveat of the train-test split approach is that it only works best when data are abundant. But in most cases, researchers might not have a large dataset. A common approach, in this case, is the $k$-fold cross-validation approach, visualized in Figure 4.1. In this approach, the full dataset is split into a fixed $k$ number of approximately equal folds, for example, 3 folds, then each fold, in turn, is used for testing, and the remainder is used for training. Typically, machine learning researchers use 10-fold cross-validation.

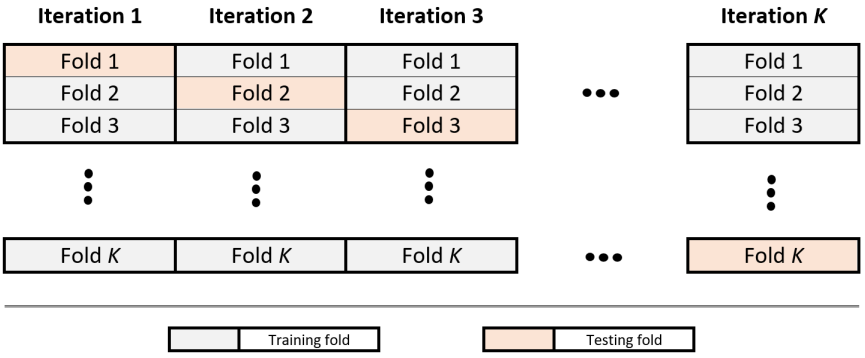



FIGURE 4.1: A Visualization of the $K$-Fold Cross-Validation Approach

There are several other validation approaches, including, for example, leave-one-out cross-validation and nested cross-validation (see Witten et al., 2016, for a comprehensive review). In this thesis, we use the standard 10-fold cross-validation approach to evaluate and compare the performance of the machine learning models.

### 4.2.3 Bias and Variance Tradeoff

The prediction error for any machine learning algorithm can be broken down into three parts: *bias error*, *variance error*, and *irreducible error*. The latter type of errors arises from unknown variables that influence the mapping of the predictors to the response variable. Therefore, regardless of what machine learning algorithm is chosen, the irreducible error cannot be reduced—hence the name. However, what can be minimized are the variance and bias errors. Variance refers "to the amount by which $\hat{f}$ would change if we estimated

it using a different training data set" (G. James et al., 2013, p. 34), whereas bias refers to the errors introduced by machine learning algorithms that approximate more complex problems through inflexible and simpler models. Technically, researchers should select a machine learning model that simultaneously achieves low variance and low bias to reduce prediction error.

## 4.3 Overall Research Design

The overall design of this thesis was made in a way that allows for transferring findings from one study to the next. While the four studies in this thesis are set to explore and address different research questions, they collectively advance one programmatic objective (*what makes videotext difficult for language learners?*). To gain tractability, in Study 1 (Chapter 6) and Study 2 (Chapter 7), I have stripped the general problem of understanding videotext difficulty to the more specific problem of investigating the independent contribution of verbal and visual complexity, respectively, to videotext difficulty. The insights gleaned from the analyses of verbal and visual complexity were used in the development of multimodal videotext difficulty models in Study 3 (Chapter 8). Finally, building upon the findings of Study 3, real-time video segment difficulty models were developed in Chapter 9.

## 4.4 A Machine Learning Pipeline for videotext Difficulty Assessment

In developing computational models for videotext difficulty, I adopted a generic machine learning pipeline commonly used in the development of supervised machine learning systems for text difficulty assessment. The pipeline involves three major stages as depicted in Figure 4.2.

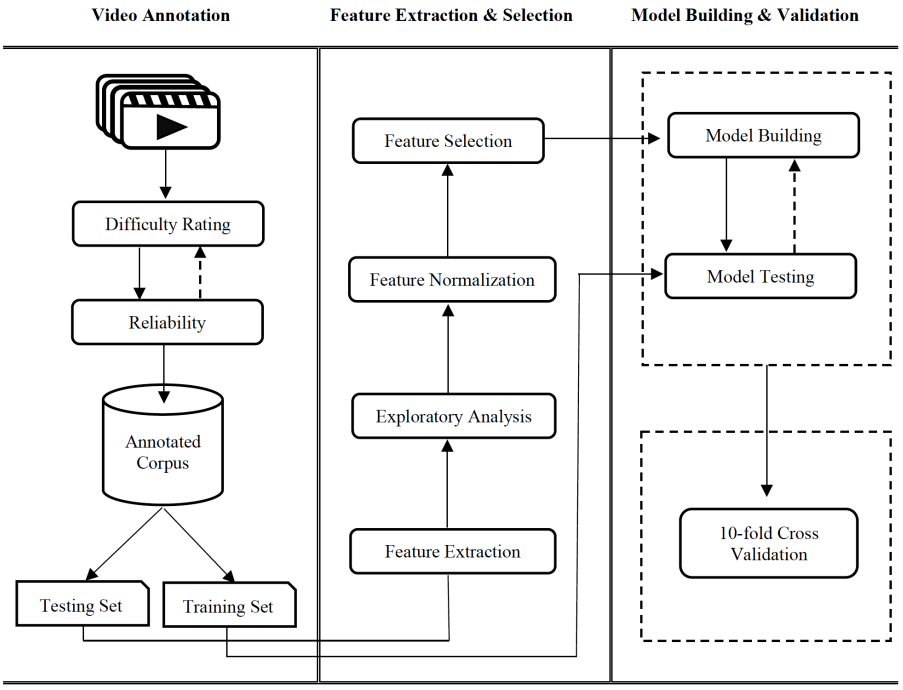



FIGURE 4.2: A Generic Machine Learning Pipeline

Each step often involves several tasks and procedures. In this chapter, I give a brief overview of these three stages, and discuss, if applicable, the tasks, analyses, and tools used at each step.

### 4.4.1 Data Collection and Annotation

After defining the goal of developing a machine system, the first task is to collect training data. For supervised machine learning problems, annotated data is required. However, the problem is data annotation is a labor-intensive and expensive process, and accordingly, there are very few annotated datasets available in the literature. This has been a true hurdle for the development of L2 text complexity models. As an alternative, researchers have used, for example, corpora of graded textbooks annotated by publishers for language learners or even using corpora of L1 texts. However, while using already annotated datasets is cheap it may not be ideal. It is very plausible that texts in these compiled datasets are graded by several publishers using different criteria. Machine learning algorithms are only as good as data they are trained on (Halevy, Norvig, & Pereira, 2009).

While there are several high-standard corpora for modeling text readability and complexity, there is, however, no corpus for videotext difficulty assessment. Therefore, I built

the Second Language Video Complexity (SLVC) corpus. The SLVC is a specialized multimodal videotext complexity corpus for developing, training, and evaluating machine learning models for videotext difficulty prediction. The corpus is composed of 640 videotexts annotated by 322 English as Foreign Language (EFL) learners for their overall complexity. The SLVC contains 320 academic lectures and 320 government advertisement video clips. These two genres were chosen because many learners are familiar with the two video genres, hence developing machine learning systems to measure their complexity will have a great utility in language learning teaching and learning. For analyzing the corpus, I adopted computational approaches and techniques from various research fields including computational linguistics, psycholinguistics, computer vision, and artificial intelligence. I describe the process of compiling and annotating the corpus in Chapter 5.

### 4.4.2 Feature Extraction

Through leveraging recent advances in computational linguistics and computer vision, I extracted a wide range of videotext features from the SLVC corpus. The feature sets cover various prosodic and phonological, lexical, syntactic, discourse, visual, and spatiotemporal complexity features, visually represented in Figure 4.3.

### 4.4.3 Data Preparation and Preprocessing

Data preparation and preprocessing is a very crucial step in any machine learning and data mining project. Real-world data is often messy, erroneous, and not in a form that is suitable to be processed by a learning algorithm. Data preparation comprises many tasks and techniques whose aim is to transform raw data into a training data set that contains no missing or incorrect values and where all features are important. Quality data helps machine learning algorithms to learn faster and generalize well on unseen data. Data preparation can take a considerable amount of processing time (Domingos, 2012; Pyle, 1999) and it involves tasks such as data cleaning, normalization, transformation, and dimensionality reduction (García, Luengo, & Herrera, 2015).

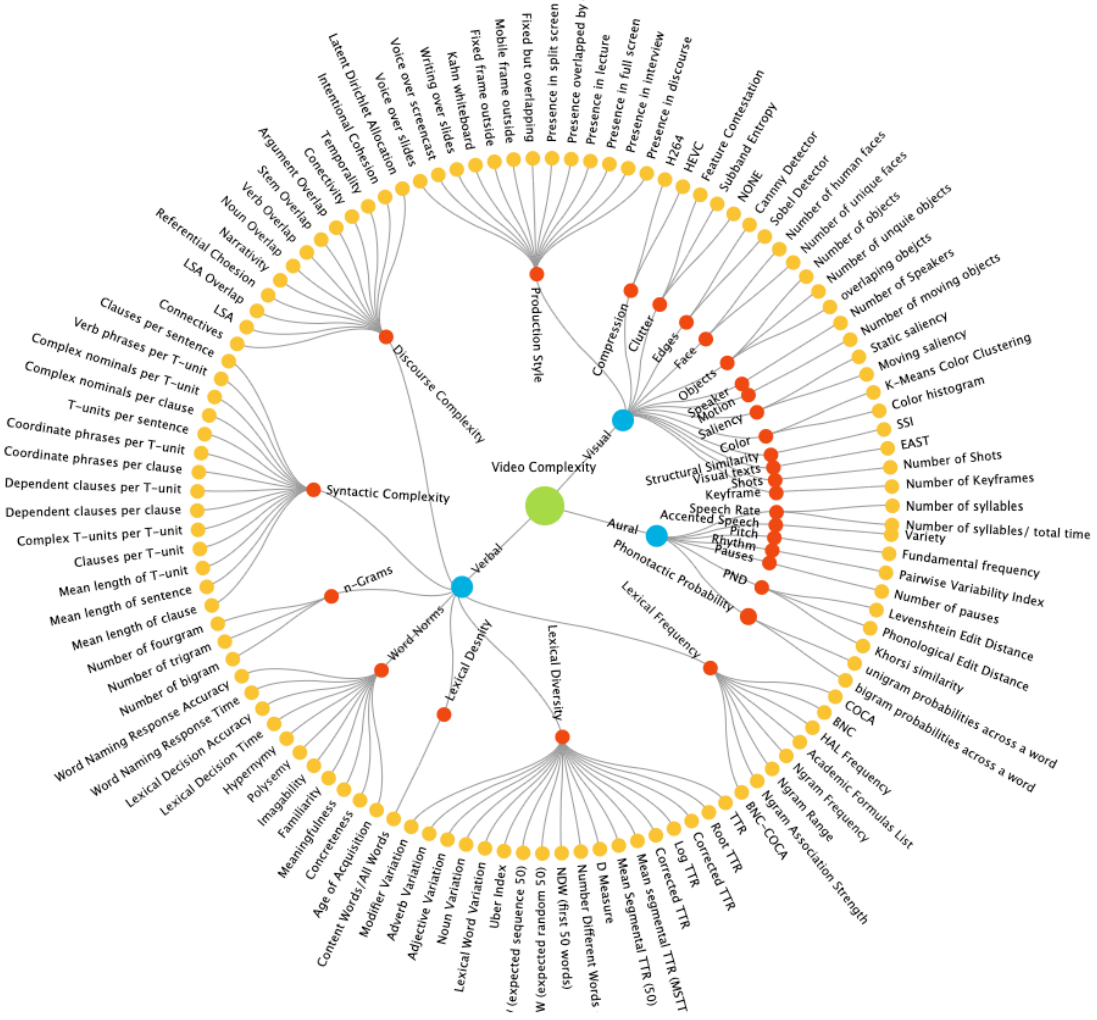



FIGURE 4.3: Verbal and Visual Complexity Features

To reduce the number of features, several feature selection—also called dimensionality reduction—techniques can be used. These methods seek to reduce the representation of the data set so that it becomes much smaller in volume, while it still maintains its integrity. That is, the reduced data set should produce the same (or almost the same) analytical results (Han, Pei, & Kamber, 2011). In doing so, they reduce the size of the problem, make the machine learning model easier to interpret, and more importantly, enhance model generalization by reducing overfitting. Generally, there are three feature selection approaches; the *filter, Wrapper*, and *embedded approach*. The filter methods use statistical testing to rank and order the importance and usefulness of independent variables to a response variable (e.g. Chi-square, ANOVA, Pearson's correlation, Mutual information). The wrapper methods create different subsets of features, each subset is used to train a learning algorithm and then tested on a hold-out set. The best subset is selected by testing the algorithm using different criteria (e.g. Forward and Backward selection). Because

wrapper methods train a new model for each subset, they are very computationally intensive. Finally, the embedded methods are a combination of two previous approaches, and they perform feature selection as part of the learning procedure. Typical algorithms of an embedded feature selection approach are decision tree-based algorithms (e.g., Random Forest) and regularized Lasso. The outcome of the process is data sets that can be readily fed into machine learning algorithms.

### 4.4.4 Model Development

The fourth step is to train a machine learning algorithm to accurately assess difficulty in videotexts. In this thesis, I experimented with different regression algorithms. My choice of algorithms was informed by similar studies in the literature and by evaluating the performances of different algorithms on the tasks.

### 4.4.5 Model Evaluation

The last step is to evaluate the trained model's performance on a new dataset that it has not been trained on. A common approach for validating the predictive ability of the trained model is the K-fold cross-validation technique (Witten et al., 2016). Depending on the machine learning algorithms choosing, the accuracy of the learner is evaluated using a test dataset. The sixth step is improving the performance of the machine learning model. This sometimes requires choosing another algorithm, utilizing more advanced strategies to improve the performance of the model, or collecting more data and doing additional preparation work on the dataset.

## 4.5 Chapter Summary

In this chapter, I discussed several key machine learning concepts pertaining to the development and evaluation of explanatory and predictive models. I also described the overall research design of the thesis and outlined the main objectives of the four studies conducted in this thesis. In the next chapter, I describe the development and annotation of the Second Language Videotext Complexity (SLVC) corpus to be used in developing computational models of videotext difficulty assessment in this thesis.

# Chapter 5

# The Second Language Videotext Complexity Corpus

In this chapter, I introduce the Second Language Videotext Complexity (SLVC) corpus, a purposefully built video corpus for investigating L2 videotext complexity and difficulty. The SLVC corpus comprises 320 academic lectures (AL) and 320 government advertisement (GA) videotexts. I first motivate the necessity and benefits of building a video corpus for researching video text complexity and difficulty in the field of applied linguistics. Then, I describe the development of the SLVC corpus and the annotation procedure. Finally, I present the results of an exploratory analysis on the SLVC corpus and the participants' ratings of videotext difficulty.

## 5.1 Making the Case for the Development of the SLVC Corpus

The availability of text corpora has brought myriad advantages to the investigation of text readability and complexity. For example, using large text corpora, readability researchers were able to investigate various dimensions of textual complexity, provide empirical evidence that validates and falsifies theoretical claims, and devise and test new complexity indices and measures. However, developing a gold-standard corpus for use in developing automated text difficulty assessment systems is expensive for individual researchers who might not be able to afford to construct large corpora. Accordingly, there are only a few

free corpora suitable for use in readability assessment research, and apparently, there is no available corpus for videotext complexity and difficulty research.

To reduce the labor cost of building a new corpus, some text difficulty researchers have built their own corpora from readily available graded textbooks, simplified versions of Wikipedia entries, or news and magazine articles written for language learners, such as the Weekly Reader corpus and the WeeBit corpus (Vajjala & Meurers, 2012) and the OneStopEnglish Corpus (Vajjala & Lucic, 2018). The readability level (or school grade) of the compiled texts and articles, decided hitherto by the original article authors or textbook publishers, is often used as ground truth in developing an automated readability tool.

In applied linguistics, extant work has been primarily devoted to investigating learners' use or production of L2, most notably through quantitative analysis of learner corpora (e.g., Granger, Gilquin, & Meunier, 2015; Le Bruyn & Paquot, 2020). There is, however, a dearth of research on videotext complexity and difficulty. This is partially due to the lack of properly annotated videotext corpora. Thereby, the development of open-accessed video corpora would facilitate video corpus research and stimulate cross-domain collaboration leading to innovative theoretical and methodological work.

With the staging increase of the amount of video data, and their current widespread use in various domains, not only in education, but also in government advertisement, health care, and business, research on video data is more needed than any time before. Similar to their utilities in written and spoken research, the development of freely accessible annotated videotext corpora can help researchers in addressing many research problems concerning L2 videotext understanding, including, for example, how second language learners perceive and deal with complexity in videotexts and how to accurately assess difficulty in videotexts, and so forth.

Against this background, I built the Second Language Videotext Complexity (henceforth SLVC) corpus to develop machine learning models for assessing and predicting videotext complexity for language learners. The SLVC is composed of 640 annotated videotexts in two video genres, academic lectures (AL) and government advertisements (GA). For the purpose of the current study, the term "academic lecture" is generic and it includes unscripted (e.g., a recording of spontaneous academic lecture) and scripted instructional videos. Government advertisements are video clips produced by official governmental bodies or agencies to inform the public about new initiatives, policies or programs. The two genres are equally represented in the corpus, 320 videos in each genre. While the SLVC corpus was principally constructed for developing and benchmarking machine

learning models for predicting L2 videotext complexity, other applications are also possible, e.g., devising and validating new measures. In the remaining of the chapter, I describe the various steps and procedures involved in the construction of the SLVC corpus.

## 5.2 Building and Annotating the SLVC Corpus

In essence, the SLVC was developed in two major stages. In the first and preliminary stage, I defined a set of criteria for searching and selecting videotexts. Then, I searched the web for government advertisements and academic lectures that meet the defined criteria and built an initial corpus of videos that satisfied all the criteria. Next, I screened the initial corpus to ensure that there were no duplicate videotexts or broken URLs. Then, I developed a script to automatically retrieve all the video metadata, and download video files, audio tracks, and transcripts into correspondence folders. In the second stage, I adapted and validated a text difficultly scale, and then I recruited a total of 322 intermediate EFL learners in a Saudi university to rate the difficulty in the corpus. In the sections that follow, I elaborate more on each of these stages in much more detail.

### 5.2.1 Videotext Selection Criteria

An essential first step in developing a high-quality corpus is defining the text selection or inclusion criteria. What determines the text selection in any corpus building project is what the corpus will be used for; that is what research questions motivate the analyses on the compiled corpus. Here the primary aim is to develop automated models for explaining and predicting in AL and GA videotexts.

At first, I developed a set of selection criteria to help in searching for and selecting videotexts to develop the SLVC corpus. The criteria were:

1. Videotexts should have no copyright restrictions.
2. Videotexts should be representative of its genre.
3. Videotexts should be about different topics.
4. Videotexts should not be about topics that are deemed taboo or culturally inappropriate for the study participants.

5. The length of videos should be short, ideally within the range of 1-12 minutes, to not overwhelm the participants and impact their judgment.
6. Videotexts should have different production styles.
7. Videotexts should not be animated in their entirety or include excessive use of animations.
8. Videotexts should not display subtitles or captions.

The selection criteria are sequentially ordered in terms of their importance; that is, the first criterion should be met before moving to the second one. So to be considered for inclusion in the corpus, a videotext should have no copyright restrictions. Content copyright restrictions are one of the obstacles in researching and developing large video corpora from the web.

### 5.2.2 Building the Initial Corpus

To build the SLVC corpus, I searched the Internet, between June 15 to August 15, 2018, for videotexts that meet the selection criteria outlined above. I adopted different searching procedures for searching for GA and AL videotexts. For GA videotexts, I visited the official websites and YouTube channels of several government agencies, departments, and offices in the United States, the United Kingdom, Australia, and Canada. When visiting each website and before selecting any videotext, I carefully read the organization's content copyright page to ascertain if there are any forms of copyright restrictions on the use of their video materials. If the videotexts are copyrighted for public use, for example, licensed under the Creative Commons Laws, I proceeded to check if videotexts meet the other remaining criteria. Once all criteria were satisfied, the URLs of videotexts were copied to a CVS file. For searching for academic video lectures, I used an advanced search filter on YouTube to find academic videos licensed under the Creative Commons Laws which had a duration of fewer than 12 minutes. I used several keywords, such as academic lecture, MOOC, biology, history, math, economy, entrepreneurship, computer science, and English lessons. Again, videos that met all the inclusion criteria were added to the initial dataset. At the end of this stage, the initial corpus contained approximately 722 videotexts.

### 5.2.3 Retrieving Video Metadata and Initial Data Screening

I wrote a python script to call the YouTube API v3 and retrieve all relevant metadata about all the videotexts in the initial corpus, e.g., video title, author, publication year, YouTube video ids, video category, and channel name. I manually checked each videotext for the appropriateness of the topic and production style. For indexing and retrieving purposes, I gave unique identifiers for all videos in the corpus that are made of two letters (GA for Government Advisement and AL for Academic Lecture) and a serial number (e.g., GA5 and AL5). After retrieving all relevant information, each video in the initial corpus has the following metadata, Video ID, URL, Title, Author, Category, Topic, Type of Copyright License, and Video Production Style. Then, I screened the initial corpus and removed duplicate entries and reviewed metadata, and added missing information. I also made sure that there is a variety of different topics and video production styles in the corpus. After screening the initial corpus, the final SLVC corpus contained a total of 640 videotexts, 320 videotexts in each video genre. Table 5.1 shows some descriptive statistics of the SLVC corpus.

Table 5.1: Descriptive Statistics of the SLVC Corpus

| Video Genre | Videos | Sentences (Mean) | Words (Mean) | Duration in minutes (Mean) |
|---|---|---|---|---|
| Government Ads | 320 | 6379 (19.99) | 103387 (324.09) | 660 (2.06) |
| Academic Lecture | 320 | 22674 (70.85) | 346054 (1081.41) | 2197 (6.86) |
| Total | 640 | 29060 (45.40) | 449616 (702.52) | 2859 (4.46) |

### 5.2.4 Retrieving Videos, Audios, and Transcripts

I developed a Python script to streamline the process of retrieving videos, audio, and video transcripts. The script calls in various extant open-sourced technologies as well as third-party commercial service to download videos, audios and phonetically transcribes the video. Given a YouTube URL, the tool downloads the video in mp4 format using Pytube library (Version 9.2.6) [1] in a local folder, then it extracts the audio in a WAV file format using the FFmpeg tool (Tomar, 2006) and saves the file into another folder. The script then makes a call to Microsoft Cognitive Service's state-of-the-art Speech-to-Text Recognition API to retrieve the phonotactic transcript of the video. After processing and cleaning the returning JSON data, the tool saves the video transcript in a plain text (txt) file in a designated folder. The three folders containing the three types of data (videos,

[1] https://pytube.readthedocs.io/en/latest/

audios, and transcripts), and are then saved into one single folder which makes the raw unprocessed version of the SLVC corpus.

## 5.3 Scale Development and Validation

Before beginning the videotext annotation sessions, I conducted a small-scale pilot study to assess the feasibility of an initial annotation plan as well as the validity and reliability of the annotation instrument. In the next section, I first present the annotation instrument, and later describe and report the findings of the validity and reliability of the modified instrument.

### 5.3.1 The Videotext Difficulty Scale

The use of the Likert-scale for annotating or rating texts difficulty is very common practice in readability and listenability research (e.g., Kotani, Ueda, Yoshimi, & Nanjo, 2014; S. Yoon et al., 2016). I adapted a questionnaire designed by S. Yoon et al. (2016) for rating the difficulty of spoken texts. The questionnaire consists of five Likert-type questions which estimate the overall comprehension of spoken text, the number of missed words, the difficulty associated with the vocabulary and acoustic dimensions (see Appendix A, for original questionnaire). To better suit the purpose of this thesis, I made some modifications to the original questionnaire questions. Firstly, I did not include the original questionnaire's first question:

Which statement best represents the level of your understanding of the passage?

1. Missed the main point
2. Missed 2 key points
3. Missed 1 key point
4. Missed 1-2 minor points
5. Understood everything

This is mainly because the question is unsuitable for the study's participants in this thesis who were at the CEFR B1 level (intermediate level) of language proficiency and hence may not be able to answer the question correctly. Secondly, I added a question (number 5) to solicit the participants' perception of the difficulty of videotext with respect to their

language ability. Thirdly, I translated the questionnaire into the participants' first language, Arabic, to avoid any misunderstanding due to language problems. The translated questionnaire was checked by a bilingual colleague and no further modifications were needed (see Appendix A.2, for the translated questionnaire).

The modified questionnaire items address different aspects of videotext difficulty. The first question asks the participants about their overall understanding of the videotext. The second question is concerned with how much information the participant still remembers after watching the videotext. The third and fourth questions address lexical and acoustic characteristics of the videotext. The last question asks the participants' overall perception of the videotext difficulty as it relates to their current language ability. The modified questionnaire is shown below:

How would you rate your understanding of the video?

1. 100%
2. 90%
3. 80%
4. 70%
5. less than 60%

How much of the information in the video can you remember?

1. 100%
2. 90%
3. 80%
4. 70%
5. less than 60%

Estimate the number of words you missed or did not understand

1. none
2. 1-2 words
3. 3-5 words
4. 6-10 words
5. more than 10 words

The speech rate was…

1. slow
2. somewhat slow

3. neither fast nor slow
4. somewhat fast
5. fast

I believe the video is…

1. much lower than my language ability
2. lower than my language ability
3. neither high nor low
4. higher than my language ability
5. much higher than my language ability

The questionnaire, however, does not include questions addressing the grammatical, syntactic, discoursal aspects of videotext. Furthermore, it did not address the visual and spatiotemporal dimensions of videotext mainly due to practical reasons. It was anticipated that addressing the multifaceted dimensions of videotext difficulty would prolong the duration of the rating session and may lead to 'respondent burden' and result in low-quality responses (Graf, 2008). Besides it was unfeasible to design questionnaire items specific to each videotext in the corpus (n=640) addressing, for example, syntactic or grammatical complexity in each videotext. Similarly, given that visual and spatiotemporal complexity in videotext are still undeveloped constructs, it was difficult to design items addressing issues related to these constructs. For these reasons, the instrument items were purposefully kept short and they only solicit the participants' overall perception of videotext difficulty.

For analyzing and scoring the participants' responses, I followed the same scoring procedure implemented in S. Yoon et al. (2016). The scale is comprised of five Likert-type questions designed to be combined into a single composite score during analysis. Each questionnaire item is scored from one to five. Therefore, the highest possible point for all five items is 25 and the lowest possible point is 5. This means that the higher the score is, the more difficult the videotext is for each participant. Since we are interested in developing a single criterion variable (henceforth videotext difficulty score), participants' responses are first aggregated and then averaged for each videotext.

It should be noted that the adopted annotation instrument is a Likert-scale and generates ordinal data; that is, the numbers assigned to different response categories express a" greater than" relationship, and the intervals between two consequent points are not always identical. Thus, certain analyses are only applicable to such types of data. But, as noted by S. Yoon et al. (2016), all questionnaire items are originally designed to measure

one single construct (perceived text difficulty), so the scale data can be treated as interval data and statistical analyses applicable to interval data, including descriptive analysis and linear regression models, can be applied on the scale data.

### 5.3.2 The Pilot Study

The purpose of the pilot study was twofold; firstly, it assessed the feasibility of an initial video annotation plan, and, secondly, it measured the validity and reliability of the annotation instrument. Concerning the annotation plan, the study sought to explore what is the most appropriate number of videos that an individual participant can rate in a single annotation session without feeling worn-out or bored. Therefore, I made four groups of videos, A, B, C, and D, each contains, 4, 6, 8, 10 videos, respectively, for the participants to rate. The videotexts in these four groups were randomly selected from the SLVC corpus, 14 AL and 14 GA videotexts. Additionally, the study aimed to estimate how many raters are needed for reliable ratings of videotexts. So, different numbers of raters (3, 4, 5, and 6) were assigned, respectively, to each group of videotexts, A, B, C, and D. Another aim of the pilot study was to test the validity and reliability of the modified and translated annotation instrument.

To this end, a total number of 24 male EFL students made the sample of the pilot study. At the time of the study, the participants were all taking an advanced English course at an English Language Institute (ELI), in a Saudi state university. The sample of the pilot study was drawn from a large population of EFL learners from which I planned to recruit more participants for my main study. I used the convenience sampling approach to select the pilot study's participants. Specifically, I made formal arrangements with the ELI's administration and the teachers' coordinator to inform and invite two intact classes of students to participate in the pilot study. I was not involved in the selection of the two classes. The teacher coordinator selected two classes and informed their instructors about the study three days before the arranged day of the pilot study. The two instructors were asked to convey key information about the study to their students, such as the purpose of the study, and that their participation in the study is not mandatory, and what task they will be asked to do if they participated.

On the day of the study, the two instructors brought in 24 students of their students (out of 55 students invited), who formerly agreed to participate in the pilot study, to a computer lab. When they were in the lab, I explained to the students, in Arabic, the purpose of the pilot study and handed them the study's consent form to read and sign. To minimize the

effect of listeners' fatigue on their ratings, the participants were informed that they can pause at any time during the session and resume whenever ready. After they returned the signed consent, I assigned each student to one of four groups based on where they were sitting. While they were annotating the videos, I took notes of their questions and remarks on the annotation process. I also monitored the time duration that each group needed to complete the questionnaire.

### 5.3.3 Findings of the Pilot Study

The analysis of the pilot study's data revealed some important findings. The first question of the pilot study was concerned with how many videotexts are appropriate for an individual participant to annotate in a single annotation session. When I asked them at the end of the annotating session, the participants suggested that watching and annotating more than five videos is more laborious and tedious and they may opt not to annotate more than 6 videotexts. Concerning how many raters are needed for reliable ratings of videotext difficulty, I performed inter-rater reliability analysis on the participants' responses in each group. The results of the Interclass Correlation (ICC) analysis showed that the participants in all four groups had low to moderate agreements, ranging from (.36 to .74).

To check the convergent validity of the instrument, I performed bivariate correlation analyses on all possible pairs among the instrument questions. The results indicated that all five questions significantly correlated with each other, with Pearson values ranging from $r = .34$ to $.80, p < .01$. The moderate-to-strong inter-correlations among the instrument items suggest that one construct is being measured: overall perception of videotext difficulty. Furthermore, to assess the reliability of the annotation instrument, Cronbach's analysis was performed on the participants' responses to the instrument items. The result showed that there is a satisfactory internal consistency among the instrument items ($\alpha$ = .81), above the suggested benchmark value of .70 (Nunnally, 1978). Post hoc analysis on the entire dataset showed that an even higher reliability score (items 2826; $\alpha$= .86).

Finally, it was observed that the number of participants who agreed to participate in the pilot study was relatively low. Therefore, I decided to use "free prize draws" as a way to induce more students to participate in the main study's annotation sessions.

TABLE 5.2: Bivariate Correlational Analysis for the Annotation Questionnaire's Convergent Validity

| Scale Items | Q1 | Q2 | Q3 | Q4 | Q5 |
|---|---|---|---|---|---|
| 1. How would you rate your understanding of the video? | - | | | | |
| 2. How much of the information in the video can you remember? | .8 | - | | | |
| 3. Estimate the number of words you missed or did not understand | .68 | .68 | - | | |
| 4. The speech rate was… | .34 | .33 | .38 | - | |
| 5. I believe the video is… | .63 | .62 | .63 | .46 | - |

## 5.4 SLVC Annotation Procedure

Building upon the results obtained from the pilot study, I made some modifications to the original annotation plan and procedure. The revised annotation procedure adopted in the development of the SLVC corpus is illustrated in Figure 5.1. Firstly, to meet the minimum advised number of videotexts (subjects) for reliable ratings, I grouped the SLVC corpus into 64 randomized groups of 10 videotexts, each of which contains an equal number of videotexts from both video genres. Each video group is further split into two sets of five videotexts (A and B) to give the participants a time break between the two annotation sessions.

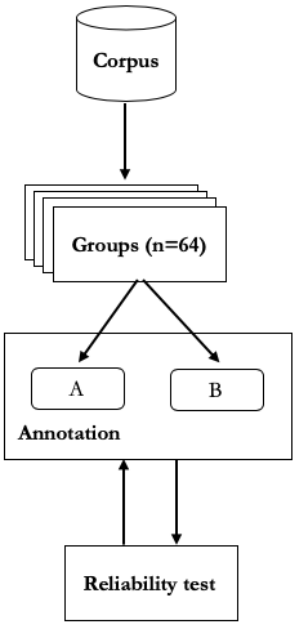



FIGURE 5.1: The Video Annotation Procedure

To ensure an acceptable level of inter-rater agreement among the raters while also human resources, I adopted an iterative process of evaluating inter-rater reliability, formally

described in equation 5.1:

$$\lambda_{ib} = \begin{cases} \text{if } ICC_i \leq .40, \gamma + 1 \\ ICC_i, \text{ otherwise} \end{cases} \tag{5.1}$$

As a starting point, three participants are assigned to each video annotation group, $\gamma = 3$. Then, at the end of the annotation session, the participants' ratings were checked for inter-rater reliability. If the ICC value was below .40, one more rater was assigned to the video group in the next annotation session.

### 5.4.1 Annotators

A total number of 322 EFL male students participated in the annotation sessions. The participants were drawn from 18 intact classes which were invited to participate in the annotation sessions. Similar to the pilot study, the class instructors initially informed their students about the study before their scheduled day. As an inducement to participate, the instructors informed their students that they would join a draw to win a valuable gift card as compensation for their time. At the time of the study, all the participants were enrolled in an intensive language preparation program. A week before enrolling in the course, all the participants took the Cambridge English placement test which assessed their language ability and placed them at the B1 level according to the CFER. The average age of the student participants was 19.4.

### 5.4.2 Annotation Setup

I used Google Forms to build a digitalized version of the questionnaire for all video groups and I made responding to all the questionnaire items required to avoid missing data. Then, I developed a simple web page to host the annotation instrument and display the videos to the participants. The annotation website and video annotation web page are showcased in Figure 5.3

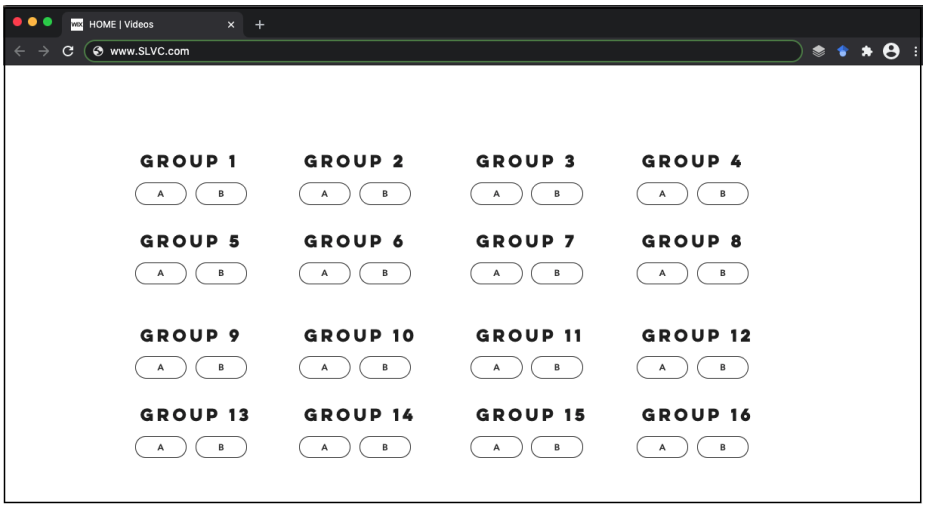



FIGURE 5.2: A Screenshot of the Videotext Annotation Website - First Page

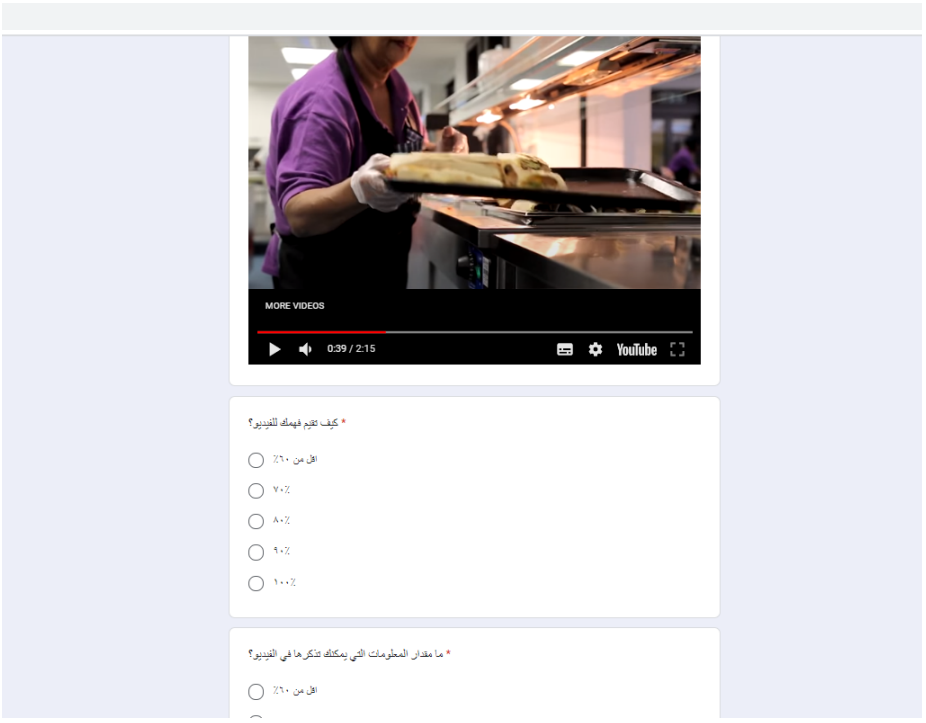



FIGURE 5.3: A Screenshot of the Videotext Annotation Website - Second Page

The annotation sessions were held in a computer lab equipped with 28 desktop computers. Before commencing the annotation sessions, I checked the lab computers, monitors, and headphones for any technical problems. I also made a shortcut of the annotation website on each computer's desktop for quick access to the annotating website in later annotation sessions.

### 5.4.3 Administering the Annotation Sessions

In total, 18 annotation sessions took place in three consecutive weeks (3 days per week, two annotation sessions per day). Each rating session lasted for approximately 60 minutes. On their designated day and time, class instructors brought in their students, who initially agreed to participate in the study, to the lab. When they were in the lab, I reminded the student participants about the main purpose of the study, explained the annotation procedures, and then handed them the plain language statement (see Appendix A.1), and the study's consent form, written in Arabic (see Appendix A.3), to read and sign.

Following this, I assigned each three students to one video group and instructed them to complete the two sets. They were also informed that they can have a time break between the video sets, pause at any time, re-watch the videos, or modify their responses if they needed to do so. The majority of the participants completed the two video sets and only 6 students opted to do only one set.

## 5.5 Analysis and Results

In this section, I report the results of inter-rater reliability analyses on all participants' ratings. Later, I present and discuss the findings of an exploratory study on the SLVC corpus in which I investigated whether video topic and genre affect the participants' ratings of videotext difficulty.

### 5.5.1 Processing and Analyzing the Questionnaire Data

After the end of each annotation session, I downloaded the Google Sheets files linked to the Google Forms (annotation form) to a local folder on my computer. Then, I replaced the "Student IDs" with anonymized identification codes (S1, S2, S3, S4, ...etc.). Secondly, I converted the participants' responses to numerical values. For example, the participants' responses to questionnaire item two were converted to numerical values as follows; "more than 10 words" = 5, "6-10 words" = 4, "3-5 words" = 3, "1-2 words" = 2, and "none" = 1. Then, I assigned a unique identification code for the processed questionnaire files and saved them in a separate folder on my computer.

Before analyzing questionnaire data, I first checked the participants' responses if they contain duplicate entries or unusual values. It was found that four participants made

duplicate entries and one participant had unusual rating values (25 points for all videotexts). The invalid responses were removed before performing any statistical analysis on the data. To make a single composite score of difficulty for each videotext, I first summed each participant's responses to the five questions (ranging from 5 to 25 points) and then averaged them. So, each videotexts in the SLVC corpus has a single difficulty score, operationalized as our criterion variable.

### 5.5.2 Inter-rater Reliability

As described earlier, an iterative approach to evaluate the reliability of the participants' ratings was adopted. Inter-rater reliability was assessed by computing the ICC for all video groups. Specifically, I adopted a two-way random model, absolute agreement, average-measures ICC (McGraw & Wong, 1996) to assess the degree of agreement among the participants' ratings of videotext difficulty. I used the R psych package to obtain the ICC results for each group after each rating session (see Appendix A.5 for the result).

In most cases, the participants rated 10 videotexts, though in four video groups the participants rated 9 videotexts due to technical problems. If the ICC value has not met the minimum ICC value of $> .40$, one or two more additional raters are recruited to rate the video group in the next rating session. Accordingly, video groups had a different number of raters, with a minimum of three raters and a maximum of nine raters. Importantly, this iterative process resulted in higher inter-rater reliability exceeding the specified threshold in most of the cases.

While the majority of video groups had an equal number of raters—that is the same participants rated videotexts in both sets (A and B), 4 video groups had an unequal number of raters due to some participants opting to rate only one set of videos instead of two. In these four cases, the *lmer* function was used as it is more appropriate for missing values.

### 5.5.3 Exploratory Analysis on Participants' Evaluation of Videotext Difficulty

I used the statistical language *R* in RStudio RStudio Team (2016) to calculate basic descriptive statistics as well as test the normality and spread of the videotext difficulty scores. The means and standard deviations for all videotexts as well as for GA and AL videotexts are shown in Table 5.3.

Table 5.3: Descriptive Statistics on the Videotext Difficulty Scores

| | Mean | SD | Median | Skewness | | Kurtosis | |
|---|---|---|---|---|---|---|---|
| | | | | Statistic | Std. Error | Statistic | Std. Error |
| Government Advertisement | 16.75 | 3.36 | 17 | -.43 | .14 | -.46 | .27 |
| Academic Lecture | 17.05 | 3.65 | 18 | -.54 | .14 | -.51 | .27 |
| All | 16.9 | 3.51 | 17.5 | -.48 | .09 | -.50 | .19 |

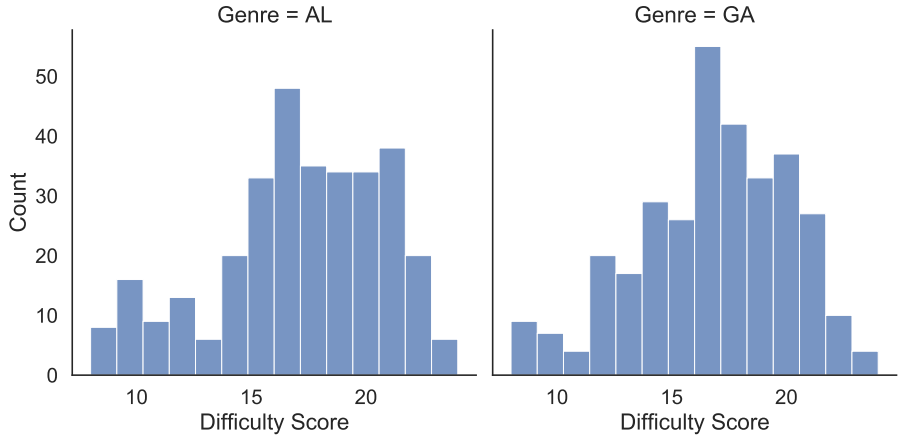



Figure 5.4: Distribution Plot of Videotext Difficulty Scores

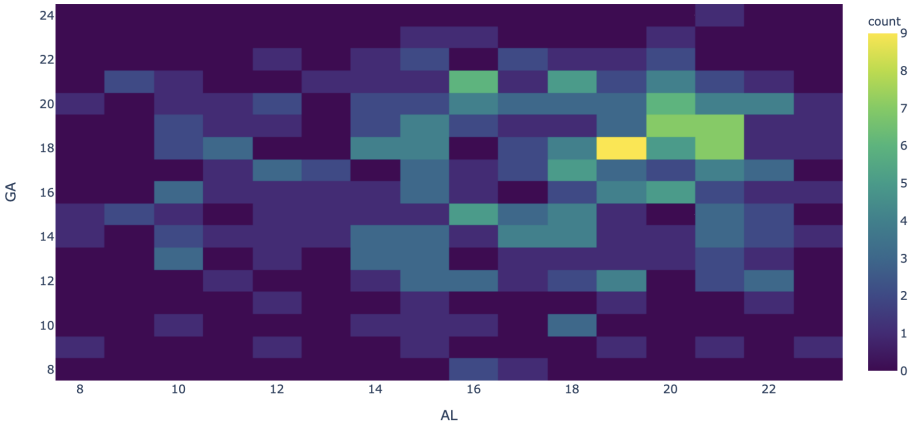



Figure 5.5: A Density Heat Map Shows Difficulty for AL and GA Videotexts

The mean and standard deviation of all videotext difficulty scores were 16.9 and 3.51, respectively. The mean of AL videotexts is slightly higher than that of GA videotexts. Outliers were visually checked using a box plot and there were no outliers found. The normality of the videotext difficulty scores was assessed numerically with a Shapiro-Wilk test and visually checked using the QQ plot (quantile-quantile plot), shown in Figure 5.6. The test resulted in a significant value, indicating the assumption of normality was not

met ($W = 0.96, p = .001$). That is, the sample distribution of data does not follow the normal Gaussian distribution, with negative skewness of -.48 (SE =.097) and kurtosis of -.5 (SE =.193).

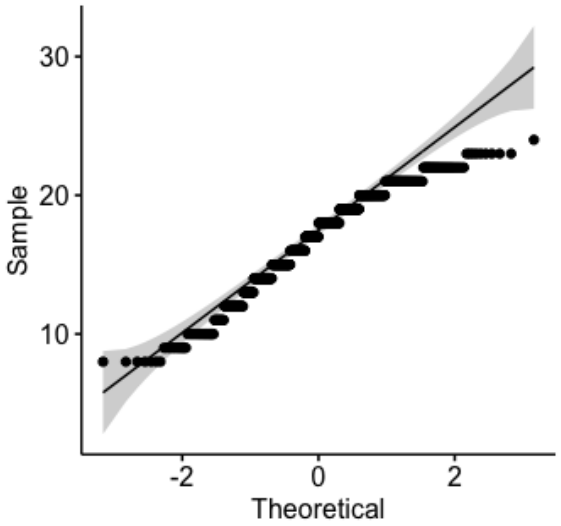



FIGURE 5.6: Quantile – Quantile Plot for Videotext Difficulty Ratings

Negative skewness indicates that the mean of the data values is less than the median, and the data distribution is left-skewed. In other words, many of the videotexts in the corpus were deemed difficult by the participants.

## 5.6 Chapter Summary

In this chapter, I described the process of developing and annotating the Second Language Videotext Complexity (SLVC). I also reported the results of validity and reliability analyses on the annotation instrument which indicated the validity and reliability of the annotation instrument. Additionally, adopting an iterative reliability assessment protocol I assessed and validated the reliability of videotext difficulty scores. Lastly, I performed an exploratory analysis on the participants' ratings of videotext difficulty. In the next chapter, I interrogate the role of verbal complexity in explaining and predicting videotext difficulty.

# Chapter 6

# The Contribution of Verbal Complexity to Videotext Difficulty

Previous research in printed and spoken texts has identified several verbal complexity features that contribute to difficulty in these types of text, however, less is known about whether these findings extend to videotext difficulty. To address this gap, this study interrogated the relative contribution of a large set of verbal complexity features on explaining the difficulty of videotext in the SLVC corpus. Using advanced computational linguistic tools, 220 verbal complexity features were extracted from the audio streams and phonetic transcripts of videos in the SLVC corpus. The predictive powers of these features were investigated, and the most predictive features were included as input in regression models to predict the difficulty of unseen videotexts. The generalizability of the regression models to predict videotext difficulty was evaluated using a 10-fold cross-validation approach. Finally, the study explored how verbal complexity varies across video genres and whether genre-specific regression models better predict difficulty in each video genre than genre-agnostic models.

Structurally, the chapter is divided into three sections. The first section outlines the verbal complexity features explored in this study and how they were extracted and computed. The second section reports on the findings of the study. Finally, the last section discusses the findings with those found in other studies in the literature.

## 6.1 Computing Verbal Complexity Features

In this study, I explored a wide range of verbal complexity features. The features set can be broadly classified into four categories: *acoustic and phonological features, lexical features, syntactic features*, and *discoursal complexity features*. Acoustic and phonological complexity features were extracted from the audio streams of the videos, whereas lexical, syntactic, and discourse features were extracted from the phonotactic transcription of the videotexts. I used Microsoft Video Indexer[1] (MVI) service to phonetically transcribe the videos. MVI employs state-of-the-art speech recognition algorithms to accurately convert audio streams into transcribed texts and automatically detect sentence boundaries. Before retrieving the transcribed data, the transcripts were checked manually and corrected using the MVI visual editor.

After retrieving video transcripts, I used the Tool for the Automatic Analysis of LExical Sophistication (TAALES, version 2.0; Kyle, Crossley, & Berger, 2018) and the Tool of Automatic Analysis of Cohesion (TAACO, version 2.0; Crossley, Kyle, & Dascalu, 2018) for measuring different indices related to discourse cohesion. I also used the official Stanford NLP python package (Qi, Dozat, Zhang, & Manning, 2019) to parse all video transcripts in the SLVC corpus and compute some measures of lexical, syntactic, and discourse complexity. Taking a running video transcript as an input, the Stanford parser segments the video transcript into a list of individual sentences and words (tokens), generates base forms of all words (lemmas), identifies their parts of speech and morphological features, and finally builds syntactic dependency relations between words in the sentences. Table 6.1 shows the types, number of verbal complexity features, and some examples of features in each category.

### 6.1.1 Acoustic and Phonological Complexity Features

#### 6.1.1.1 Speaker Accent

For determining what English variety or accent is spoken in the videotexts, I categorized the accents of speakers based on two factors: their countries of origin and their institution's whereabouts. For example, in the case of government advertisements, videos that were produced by an American, Australian, British, and Canadian government office or agency is presumed to have speakers who speak American, Australian, Canadian,

[1] https://azure.microsoft.com/en-au/services/media-services/video-indexer/

and British accent, respectively. The same procedure was also used in determining the speaker's accent in the academic lecturer videos, though in some videos (Khan academy videos) information about the speaker's country of origin was not easily accessible. In this case, I relied on my intuition to judge the speaker's accent as belonging to one of the four common varieties of English: American, Australian, Canadian, and British English. It is understood that each of the four English varieties is spoken in multiple accents and local dialects, but these dialectal differences were ignored in this research.

#### 6.1.1.2 Speed, Phonation Time, and Pauses

The delivery of the speech was assessed through three measures: speech rate, articulation rate, and average syllable duration (ASD). Speech rate was expressed as the number of syllables per second including pauses (number of syllables/phonation time), whereas articulation rate was operationalized as the number of syllables per second excluding pauses (number of syllables/total time). ASD is the phonation time divided by the number of syllables. Phonation time is computed as the ratio of talking in seconds. Pauses are salience periods exceeding .25s (e.g., Révész & Brunfaut, 2013). All the indices were computed in PRAAT software (Boersma & Weenink, 2019) using a script written by (de Jong & Wempe, 2009).

#### 6.1.1.3 Phonotactic Probability and Phonological Neighborhood Density

First, the orthographic form of each word token was transformed to a phonetic form using the Phonological CorpusTools (PCT version 1.3; Hall et al., 2017). The PCT makes it possible to look up each token in the Irvine Phonotactic Online Dictionary (IPhOD; Vaden, Halpin, & Hickok, 2009). The IPhOD corpus is a large collection of phonetically transcribed English words and pseudowords that were compiled at the University of California, Irvine, for research on speech perception and production (Vaden et al., 2009). Then, the phonotactic probability was calculated using average unigram or bigram positional probabilities across a word in a videotext (Vitevitch & Luce, 2004). For a word like a blink in English, the unigram average would include the probability of /b/ occurring in the first position of a word, the probability of /l/ in the second position, the probability of /i/ occurring in the third position, the probability of /n/ occurring in the fourth position of a word, and the probability of /k/ occurring in the fifth position of a word. Each positional probability is calculated by summing the log token frequency of words containing that

segment in that position divided by the sum of the log token frequency of all words that have that position in their phonetic transcription.

Phonological neighborhood density was calculated using the Levenshtein edit distance algorithm. For each query word, PCT checks each other word in the corpus for its similarity to the query word and then adds it to a list of neighbors if sufficiently similar.

#### 6.1.1.4 Rhythm and Pitch.

The rhythm was assessed in terms of the vocalic Pairwise Variability Index (PVI), which measures the extent to which one vocalic interval is different from its neighbor in duration. The normalized PVI was computed using Equation 6.1

$$nPVI = 100 \left[ \sum_{k=1}^{m-1} \left| \frac{d_k + d_k + 1}{(d_k + d_k + 1)/2} \right| / (m - 1) \right] \tag{6.1}$$

Pitch was calculated using a measure of the fundamental frequency (F0), which estimates vocal cord vibrations per second in voiced sounds. The mean number of minimum and maximum pitch, and the standard deviation were calculated for each audio stream.

### 6.1.2 Lexical Complexity

The lexical complexity of the videotext was assessed through various measures of lexical frequency, sophistication, diversity, density, psycholinguistic norms, and n-gram strength of association.

#### 6.1.2.1 Lexical Frequency

The lexical frequency was assessed using a band-based approach (e.g., Laufer & Nation, 1995; Morris & Cobb, 2004) because it is more appropriate for receptive lexical frequency analysis and more interpretable than count-based approach (Crossley et al., 2013). Frequency indices were based on lemmatized BNC-COCA word family lists. Specifically, a python script was developed to calculate the proportions and percentages of all words (AW), function words (FW), and content words (CW) belonging to the most frequent word families in the K1, K2, K3, and K4 bands, and those were not these lists (off-list words).

Further, the percentages of words included in the 2,000-frequency band (K1 + K2 words) and the 3,000 frequency of the BNC-COCA lists were also computed. While FW is closed-class grammatical words (e.g., preposition), there is notable variability in how CW was defined in previous studies (e.g., O'loughlin, 1995; Ure, 1971a). In this study, CWs refer to "nouns, adjectives, verbs (excluding modal verbs, auxiliary verbs, "be," and "have"), and adverbs with an adjectival base, including those that can function as both an adjective and adverb (e.g., "fast") and those formed by attaching the –ly suffix to an adjectival root (e.g., "particularly")" (Lu, 2011a, p. 192).

In addition to general word frequency, I calculated the proportion of formulaic sequences appearing in specialized vocabulary lists; the Academic Word List (AWL Coxhead, 2000), the Academic Formulas List (AFL Simpson-Vlach & Ellis, 2010), and Academic Spoken Word List (ASWL Dang, Coxhead, & Webb, 2017). The AWL comprises 570 academic word families which were compiled from a corpus of 3.5 million running words of written academic journal articles and textbook chapters from a total of 28 subject areas in four broad disciplines. Simpson-Vlach and Ellis (2010) developed the Academic Formulas List (AFL) which comprises multiword units that are pervasive in academic language.

TAALES was used to calculate the frequency of tokens in a text that appears in these lists and then dividing that numbers by the number of words in the text. Finally, I used the proportion of videotext words appearing in the English Vocabulary Profile (EVP) [2] (English Profile, n.d.) as an indicator of lexical complexity (Xia Wu, 2016). The EVP is an online tool that offers reliable information about which words, phrases, and idioms are known by English language learners at each level of the Common European Framework (CEF). The EVP is based on research using the Cambridge Learner Corpus and it contains several hundred thousand examination scripts written by English learners from all over the world.

#### 6.1.2.2 Lexical Sophistication

Lexical sophistication was operationalized as words, nouns, and verbs that are not in the first 2000 most frequent BNC-COCA lists of lemmatized word families.

[2] https://www.englishprofile.org/wordlists

#### 6.1.2.3 Psycholinguistic Norms

Several psycholinguistic properties of words have been of interest to researchers in cognitive science, psycholinguistics, and second language acquisition. Of particular interest to text complexity researchers have been the psycholinguistic properties of concreteness, familiarity, imageability, meaningfulness, and age of acquisition (AoA). These indices are generally calculated based on information retrieved from large databases such as the MRC Psycholinguistic Database (Coltheart, 1981), which contains a collection of human ratings of more than 150,000 words on 26 psycholinguistic norms.

I used TAALES to calculate several indices related to information derived from the MRC database (Brysbaert et al., 2014; Coltheart, 1981). TAALES also calculates indices related to human ratings of the age at which a particulate word is learned, based on the age-of-acquisition AoA list (Kuperman et al., 2012). The AoA lists contain 30,121 English content words (nouns, verbs, and adjectives) and their values which have been found to positively correlate with both response times and accuracy scores on lexical decision tasks (Kuperman et al., 2012).

#### 6.1.2.4 N-gram Strength of Association

Association measures of n-grams assess the degree to which words in n-grams occur together. Two well-attested association measures are Mutual Information (MI) and t-score (Church & Hanks, 1990; Hunston, 2002). Though the two measures compare how often an n-gram (collocations) appears in a corpus with how often it would be predicted to appear based on the frequency of the words that compose it, MI tends to highlight n-grams made up of low-frequency words whereas t-score highlights those made up of high-frequency words (Bestgen & Granger, 2014). Other approaches to word association strength are based on the directionality of the association, that is quantifying whether word1 is more predictive of word2 or the other way around (Gries, 2013). A well-established directional measure is ΔP (Delta P) which calculates the probability of word occurrence based on a cue (another word). The score is calculated using the formula: $P(O|C)\check{}P(O|\check{}C)$; that is, delta P is the probability of an outcome given a cue minus the probability of an outcome without the cue (Kyle et al., 2018).

Table 6.1: Number of Verbal Complexity Features and their Respective Categories

| Category | Subcategory | Features(n) | Example |
|---|---|---|---|
| Acoustic & phonological | | 14 | Speech rate, Pitch, Phonetic probability |
| Lexical | Frequency | 37 | Frequency of word in K1, K2, K3 and K1-4 |
| | Sophistication | 4 | |
| | Psycholinguistic | 17 | Concreteness, AOA |
| | N-gram | 27 | Frequency |
| | Density | 2 | |
| | Diversity | 40 | Log TTR, D, |
| Syntactic | | 14 | Parse tree depth, Complex nominal per clause |
| Discourse | Cohesion | 29 | Connectivity, LDA, Argument overlap |

#### 6.1.2.5 Lexical Density

Lexical density was operationalized as the proportion of content words (CWs) to running words in the video transcripts.

#### 6.1.2.6 Lexical Diversity

Indices of lexical diversity gauge the variety of unique words (types) in a text in relation to the total number of words (tokens). A conventional method to measure lexical diversity is the type-token ratio (TRR). The TRR index is based on a simple ratio of types to all tokens in a text. One obvious problem of TRR and all indices based on word frequency is that their values are affected by text length; that is as the number of tokens increases the likelihood of those tokens being unique is low. As an alternative method, transformants of the TTR, e.g., Root TRR 6.2, Corrected TRR 6.3, and logarithmic TRR, are often used which can reduce the text length problem.

$$Root\ TTR = \frac{Total\ Number\ of\ Types}{\sqrt{Total\ Number\ of\ Words}} \tag{6.2}$$

$$Corrected\ TTR = \frac{Total\ Number\ of\ Types}{\sqrt{2 \times\ Total\ Number\ of\ Words}} \tag{6.3}$$

In the past decade, several approaches to assess lexical diversity have been proposed, each purport to measure lexical diversity while having little or no sensitivity for text length. Examples of such measures are MLTD and vocd-D which use estimation algorithms (McCarthy & Jarvis, 2010). Covington and McFall (2010) proposed the moving-average type-token ratio (MATTR), which comprises of the average of TTR for overlapping $n$ words sequences in a text (e.g., 1-50, 2-51, 3-52).

In a recent study, Treffers-Daller et al. (2018) found traditional measures of lexical diversity (e.g., TTR and the Index of Guiraud) more effective in discriminating between texts of different CEFR levels than more sophisticated measures (e.g., D, HD-D, and MLTD) provided text length is kept constant across texts.

As these indices gauge lexical diversity in different ways, researchers, such as McCarthy and Jarvis (2010), recommended the use of multiple indices instead of a single index which may not capture all aspects of lexical diversity. Given that and also these measures have not been explored for videotext difficulty assessment, I decided to use both traditional (e.g., Root TRR) and more recent measures of lexical diversity (e.g., D, HD-D, and MLTD) for all words, content, and function words.

### 6.1.3 Syntactic and Grammatical Complexity Features

Before analyzing spoken data, one needs to decide on how to divide transcribed data into units in order to assess features such as syntactic and grammatical complexity. There are various units in use in the literature and several definitions and guidelines on what they are and how they can be found in the spoken sample. While there is a limited set of options for applied linguists working on written data, there is a variety of choices for spoken text analysis, including segmenting spoken data based on some predefined semantic, syntactic, and intonation criteria (Foster, Tonkyn, & Wigglesworth, 2000). More often than not, the choice of analysis or production unit depends on the nature of the data (formal vs. informal, monologic vs. dialogic). Among several units of analysis, T-unit and AS-unit are commonly used in spoken research, with the latter being more suitable for the analysis of conversational (informal) spoken language. T-unit is essentially a main clause plus all subordinate clause attached to it (Hunt, 1965, 1970), whereas AS-unit is "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (Foster et al., 2000).

In this study, the video transcripts were segmented into T-units, as opposed to AS-units or C-units, for three reasons. First, both academic lecture and government advertisements videotexts in the SLVC corpus are carefully scripted monologues speech and contained minimum occurrences of false starts, repetitions, and self-correction, hence, they seem to lean towards exemplifying formal speech rather than informal or conversational speech. The second reason for this decision is based on the findings of a corpus study by Biber, Gray, and Poonpon (2011) who found complexity metrics based on T-unit and dependent clauses are more suitable for the analysis of spoken text rather than written text. Besides,

Halliday (1989) theoretically contended that major grammatical complexities of spoken language involve dependent clauses. The third reason is merely practical. Segmenting spoken data into AS-units requires identifying points in the speech stream that can accurately and reliably segment speakers' utterances.

The challenge of choosing the syntactic or production unit notwithstanding, in the realm of L2 research, a large variety of indices has been employed to operationalize the construct of syntactic complexity. This even makes it more difficult for language researchers to choose the right complexity metric for their structure of interest. These indices may not necessarily tap into different aspects or dimensions of syntactic complexity in text, however. As noted by Norris and Ortega (2009) some complexity indices are redundant when used together because they measure the same construct, and conversely, other indices are distinct and contemporary, and thereby they should be used and interpreted together because they index different dimensions of complexity. Using redundant and correlated measures in a single study is thereby pointless, and it even can lead to psychometric issues and unreliable findings (Norris & Ortega, 2009).

To extract syntactic and grammatical complexity from a large corpus, one needs to identify the parts of speech (POS) of the words in a sentence, the syntactic (i.e., constituency), and/or grammatical relationships (dependency) between them. Recent advances in computational linguistics have made the automatic extraction of such textual features possible, while the improvement in syntactic parsers has made their outputs more accurate and reliable. Syntactic parsers generally require the text input to be segmented into individual sentences before they are tokenized and part-of-speech (POS) tagged. While some videotexts in the corpus had already their transcripts professionally edited and polished by their original providers (i.e., government agencies or education institutions), there were also some videotexts in the corpus that lacked proper caption. Therefore, I used Microsoft's Azure speech-to-text recognition technology to phonetically transcribe the videotexts. Microsoft Azure employs state-of-the-art algorithms to accurately convert audio streams into transcribed texts and detect automatically sentence boundaries. Before retrieving the transcribed data, the transcripts were checked manually and corrected using the MVI's visual editor.

Once all the video transcripts were edited, retrieved, and preprocessed, I used the official Stanford NLP python package (Qi et al., 2019) to parse all video transcripts in the SLVC corpus. Taking a running video transcript as an input, the parser segments the video transcript into a list of individual sentences and words (tokens) and generates base forms of all words (lemmas), their parts of speech, morphological features, and build a syntactic

dependency between words in the sentences. Additionally, I used Berkeley's constituency neural parser (Kitaev & Klein, 2018) which outputs a hierarchical representation of phrase structure rules (often called parse tree). The two parsers generate valuable information about text sentences which can be used for calculating different syntactic and grammatical complexity indices of various levels of granularity. Constituency parsers, for example, are essential for identifying phrasal and clausal boundaries and helps to differentiate independent clauses and subordinate clauses. For example, the parse tree of the short sentence,*"This time around, they're moving even faster"*, is (S (NP (DT This) (NN time) (RP around)) (NP (PRP they)) (VP (VBP 're) (VP (VBG moving) (ADVP (RB even) (RBR faster))))), which can be visually represented as in Figure[3] 6.1:

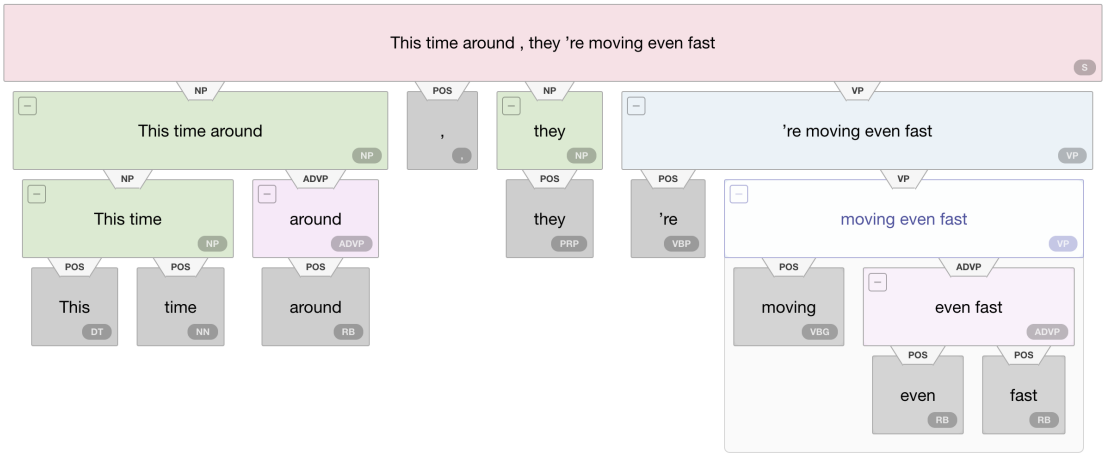



FIGURE 6.1: A Visual Representation of Parse Tree for the Sentence: *This time around, they're moving even faster*

Dependency parsers, on the other hand, analyze the grammatical structure of a sentence and generate binary grammatical relationships or dependency relations between lexical items in the sentence. Specifically, the parser outputs a collection of dependency relations between each headword (or governor) and words that modify them (their dependents). For our example sentence, This time around, they're moving even faster, the Stanford neural dependency parser outputs the following dependency relations:

The dependency relations of the sentence are visualized in Figure 6.2. Taking a single dependency relation as a way of example, the word faster modifies moving and the nature of dependency is *advmod* (adverb modifier). Note that the main verb (moving) is the dependent of the ROOT of the sentence and that single lexical items can be represented both as governors and dependents in different grammatical relations (e.g., the word time

[3]The Figure was produced by AllenAI parser, https://demo.allennlp.org/constituency-parsing, for demonstration only

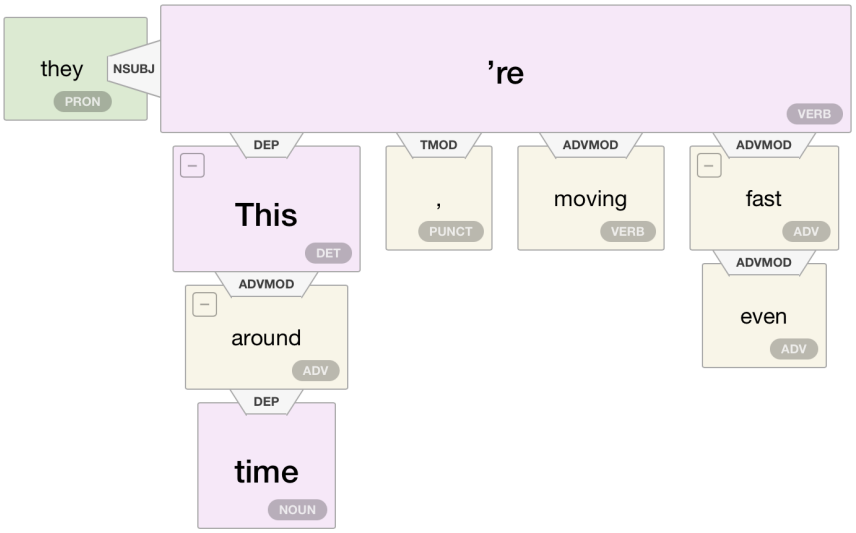



FIGURE 6.2: Graphic Representation of Dependency Parse for the Sentence: *This time around, they're moving even faster*

is both a governor of the and a dependent of moving). Dependency parsers differ in their outputs based on theoretical and practical choices made by their developers. They also differ in their tag convention. For example, while the RASP dependency parser (Briscoe, Carroll, & Watson, 2006) tags 17 grammatical relations, the Stanford Dependency parser includes 50 grammatical relations (D. Chen & Manning, 2015; Manning, Bauer, Finkel, & Bethard, 2014).

Using the syntactically-parsed video transcripts, I computed 14 syntactic complexity indices proposed by Lu (2011b). These indices measure the length of a structure (clause, T-unit, and sentence), amount of subordination, amount of coordination, degree of phrasal sophistication, and overall sentence complexity. Lu (2017) argues that these indices align with the four dimensions of syntactic complexity suggested by Norris and Ortega (2009), and hence they can be expected to capture syntactic variety and sophistication. Also, the number of negations in a videotext is counted and then normalized by the total number of sentences in the videotext.

In addition to these indices, I computed the mean dependency distance (MMD) (H. Liu, 2007; H. Liu, Xu, & Liang, 2017) in the videotext. Dependency distance is defined as "the number of words intervening between two syntactically related words, or their linear position difference in sentence indices based on dependency distance"(H. Liu et al., 2017, p.172). For example, the dependencies between words in Figure 6.2 can be adjacent or nonadjacent, that is the two words forming a dependency may come next to each other

or be separated by intervening words. Dependency distance is generally held as an important index of memory burden (i.e., Dependency Locality Theory (Gibson, 1998, 2000) and can be used as an indicator of syntactic difficulty. According to the Dependency Locality Theory, the syntactic complexity of a sentence can be predicted by two factors: the "storage cost" of maintaining the previous words in memory and the "integration cost" of connecting the words to previous words in memory. Consequently, the greater the dependency distance, the heavier the memory load the word places on the speaker or hearer's mind. H. Liu (2007) proposes the following two formulas to calculate the mean of dependency distance of a sentence and a text:

$$MMD(Sentence) = \frac{1}{n-1}\sum_{i=1}^{n}|DD_i| \tag{6.4}$$

where $n$ is the number of words in the sentence and $DDi$ is the dependency distance of the $i$th syntactic link of the sentence, and

$$MMD(Text) = \frac{1}{n-s}\sum_{i=1}^{n}|DD_i| \tag{6.5}$$

where $n$ is the total number of words in the text, and $s$ is the total number of sentences in the text. The output of the Stanford dependency parser generates the required information to calculate MMD.

### 6.1.4 Discourse Cohesion Features

Among the various discoursal complexity features, in this study, I examined whether discourse cohesion was a good predictor of video lecture difficulty. Cohesion can be considered at the local (sentences), global (paragraphs), and text levels. Because paragraph boundaries in the phonetic transcripts of the videos cannot be easily identified, we only considered local and text-level cohesion in this study.

I used TAACO, version 2.0 (Crossley, Kyle, & Dascalu, 2019) to measure local and overall cohesion in video transcripts. In particular, I analyzed the video transcripts in terms of various types of connectives and lexical overlap across sentences and throughout the video transcripts. Connectives can be classified by the type of cohesion they create additive, causal, logical, or temporal (Halliday & Matthiessen, 2014) and whether they extend the ideas described in the text (positive connectives) or not (negative connectives

(Sanders, Spooren, & Noordman, 1992). Using TAACO, we computed the ratio of additive (e.g. further, in sum), causal (e.g. because, nevertheless), logical (e.g. if, consequently, therefore), and temporal (e.g. when, after, until) connectives to the total number of words in video transcripts. We also employed TAACO to calculate the average scores for all lemmas and content lemmas that overlap between two or three adjacent sentences as well as the frequency of sentences that include overlapping lemmas.

Local cohesion refers to text features that link short text segments such as sentences together, while global cohesion refers to text features that link larger segments of texts (i.e., paragraphs). Another distinction between local and global cohesion is that global cohesion indices tap into coherence while local cohesion indices do not (Crossley et al., 2016).

Givenness indices approximate the proportion of given information to new information by calculating overall pronouns density, pronouns to nouns ratio, repeated content lemmas, and pronouns. Furthermore, TAACO computes several indices related to semantic overlap between text segments using sophisticated computational models used in NLP research, such as Latent Semantic Analysis (LSA; Landauer et al., 1998), Latent Dirichlet Allocation (LDA; Blei et al., 2003), and word2vec vector space representations (Mikolov, Chen, Corrado, & Dean, 2013). For each of these models (LSA, LDA, word2vec), TAACO calculates the average similarity between progressive adjacent text segments (sentences) in the text.

## 6.2 Statistical Procedure and Analysis

### 6.2.1 Research Questions

This study in this chapter addresses the following research questions:

1. What is the relationship between verbal complexity and language learners' judgments of videotext difficulty?
2. What verbal complexity features are the most predictive of video lecture difficulty?
3. To what extent can we predict videotext difficulty using measures of verbal complexity?

4. Do verbal complexity features differ between academic lecture and government advertisement videotext? And if so, what are the most predictive verbal complexity features of videotext difficulty in each video genre?

### 6.2.2 Corpus

In this study, I used the entire SLVC corpus for developing genre-agnostic/independent machine learning models, whereas I used the two subcorpora, academic lecture (AL) and government advertisement (GA), in building genre-specific models for each video genre. I use the "SLVC corpus" or simply the corpus to refer to the two types of videotexts when no confusion will result.

### 6.2.3 Explanatory Analysis and Data Pre-processing

After extracting all verbal complexity features under investigation, some preliminary prepossessing procedures were applied to the dataset before running machine learning modeling on the corpus. First, the categorical feature (speaker accent) was converted to numerical values using the One Hot Encoding technique. Secondly, the entire dataset was shuffled and divided into a training set and test set following the 67/33 split approach (Witten et al., 2016). This resulted in a training set of 512 videotext samples and a testing set of 128 videotext samples. The same procedures were applied to AL and GA subcorpora, which resulted in 256 videotexts for training and 64 videotexts for testing for each video genre.

Initial inspection on the features identified that the values of some features were approximately similar across observations. To address these issues, low-variance features were dropped because they do not have useful information. In most cases, discarded features represented syntactic features that occurred extremely rarely in spoken texts. Then, because our features were not on the same scale, all numeric features were scaled using the standard Scaler formula implemented in Scikit-Learn (Pedregosa et al., 2012). The formula standardizes features by removing the mean and scaling to unit variance using the following equation:

$$Z = \frac{x - \mu}{\sigma} \tag{6.6}$$

The same preprocessing routine was also applied to the testing sets.

Then, I performed correlational analyses between these features and the participants' rating of difficulty for AL, GA, and all videotexts. A correlational analysis is useful in beginning to pinpoint what verbal complexity features have a colinear relationship with the videotext difficulty. To reduce the number of features, I adopted an embedded approach for feature reduction regularized linear regression model, LASSO (short for Least Absolute Shrinkage and Selection Operator algorithm). Lasso penalizes the regression coefficients with an L1 penalty so that the coefficients of 'less important' features shrink and become zeros (Tibshirani, 1996). Therefore, any features which have non-zero regression coefficients are selected by the LASSO algorithm as possibly meaningful predictors. The LASSO alpha coefficient was tuned during a 10-fold cross-validation process and the best model was used to select verbal complexity features. Next, using the selected features I developed several regression algorithms, including linear Support Vector Machine, Elastic Net regressor, Decision Tree Regressor, and Ridge regressor, Ordinary Least Square (OLS) regressor. A comparison between the performances of these models on predicting videotext difficulty on the testing set showed that a simple OLS regression model is on par with more complex algorithms. Therefore, I decided to use OLS regression to model videotext difficulty because of its interpretability.

After developing the model on the training set, coefficients were checked for both variance inflation factors (VIF) values and tolerance to inspect if the model suffers from multicollinearity and to avoid overfitting. Additionally, if any of the features were not significant, they were removed, and the remaining features were fitted again into a regression model. The results of this analysis were later extended using the regression model to the held back, independent test set.

To compare our model performance, I developed a baseline model using the most widely used readability index, the Flesch-Kincaid reading ease formula (Flesch, 1949). The formula takes the average number of words per sentence as the indication of syntactic difficulty and the average number of syllables per word as the estimation of word complexity.

To further check the generalizability of the regression model, the final model was evaluated using a 10-fold cross-validation approach. In this validation approach, the entire dataset is divided into 10 folds and nine of these folds are used to develop a regression model that is then tested on the left-out fold. This process is repeated 10 times, so all data are used to both train and test the model. I selected this approach because several experiments have shown it is a more reliable approach for emulating machine learning models performance (Witten et al., 2016).

The performances of the models were compared using three general evaluation metrics, Pearson's correlation coefficient, adjusted R squared, and Root Mean Squared Error (RMSE). For performing descriptive and correlational statistical, I used the RStudio software (RStudio Team, 2016) and used the Scikit-Learn library for building and evaluating machine learning models (Pedregosa et al., 2012).

## 6.3 Results

### 6.3.1 Correlation between Videotext Difficulty and Measures of Verbal Complexity

To examine the relationship between any of the verbal complexity indices and language learners' rating of videotext difficulty, a Pearson's correlation analysis was performed on standardized verbal complexity features. The results showed 79 verbal complexity features significantly correlated with the participants' ratings of videotext difficulty ($p <$ .001) and above $r = .15$ (i.e., a small effect size (Cohen, 1988). Table 6.2 shows the ten most highly correlated verbal complexity features with ratings of videotext difficulty along with their means and standard deviations. The correlation coefficients of the remaining indices can be found in Appendix A.8.

To further explore the interaction between verbal complexity and video genres, measures of complexity of videos in the two genres were correlated with participants' ratings of videotext difficulty. The results showed statistically significant correlations between 125 complexity features and the difficulty of AL videotexts (see Table 6.3, for the highly correlated features). The correlation analysis on GA videos showed that only 47 verbal complexity features had collinear relationships with the participants' ratings of the difficulty of this video genre. Table 6.4 shows highly correlated verbal complexity features with participants ratings of difficulty in GA videotexts.

### 6.3.2 Feature Subset Selection

Correlational analysis showed that several verbal complexity features (i.e., 125 features for AL subcorpus) have significant collinear relationships with videotext difficulty. For this study, I explored two embedded methods, Random Forest and Lasso, for feature selection.

Table 6.2: Highly Correlated Verbal Complexity Indices with The Participants' Ratings Of Videotext Difficulty (All Corpus)

| Category | Feature | *M* | *SD* | *r* |
|---|---|---|---|---|
| Acoustic and phonological | Phonation time | 224.15 | 187.7 | .35 |
| | Frequency of syllables | 1033.75 | 932.75 | .34 |
| | Frequency of pauses | 57.31 | 62.5 | .22 |
| | Pitch (SD) | 66.44 | 21.92 | -.28 |
| Lexical | Root TTR (AW) | 8.28 | 2.21 | .42 |
| | Age of Acquisition (CW) | 6.18 | 0.61 | .41 |
| | K3 (AW) | 6.65 | 3.73 | .40 |
| | K1-2 (AW) | 83.15 | 7.03 | -.37 |
| Syntactic and grammatical | Complex nominals per clause | 1.04 | 0.37 | .37 |
| | Mean length of clause | 9.39 | 2.28 | .31 |
| | Complex nominals per T-unit | 2.09 | 1.36 | .20 |
| | Dependent clauses per clause | 0.40 | 0.11 | .16 |
| Discourse | Coordinating Conjuncts | 0.009 | 0.008 | -.25 |
| | All causal (normalized) | 0.016 | 0.010 | -.21 |
| | Pronoun to noun ratio | 0.25 | 0.14 | -.18 |
| | Positive causal (normalized) | 0.023 | 0.012 | -.18 |

*Note.* All correlations are significant at the .01 level (2-tailed)

Features selected by Lasso yielded better results. The lasso lambda was determined using a 10-fold cross-validation approach and the best model selected 29 features for AL videos, 16 features for GA videos, and 37 features for all videos.

### 6.3.3 Regression Models

Multiple regression analysis was conducted to explore if the selected verbal complexity features significantly explain the variance in the participants' ratings of videotext difficulty. After removing insignificant features (e.g., with coefficients exceeding the common alpha level of 0.05), the final model yielded a significant model, $F(9, 418) = 29.47, p < .01, r = .62, R^2 = .38$ for the training set. The model includes 8 verbal complexity features: *standard deviation of the pitch, K1-4 (CW), AWL Sublist-3, AFL, ASWL Sublist-2, Root RTT (FW), complex nominals per clause, age of acquisition (CW)*, and *COCA spoken bigram AC*. These features tap into the acoustical, syntactic, and lexical dimensions of videotext complexity and they jointly explained 38% of the variance in the participants' judgments of videotext difficulty. Table 6.5 shows additional information regarding this model and the beta weights for each index in the training model.

Table 6.3: Highly Correlated Verbal Complexity Indices with the Participants' Ratings of AL Videotext Difficulty

| Category | Feature | *M* | *SD* | *r* |
|---|---|---|---|---|
| Acoustic and phonological | Phonation time | 338.29 | 200.51 | .52 |
| | Frequency of pauses | 98.71 | 62.92 | .36 |
| | Pitch (SD) | 65.64 | 14.88 | -.34 |
| | Speech rate | 3.68 | 0.65 | .32 |
| Lexical sophistication and diversity | K3 (AW) | 7.45 | 3.76 | .64 |
| | K3 (CW) | 68.26 | 10.32 | .63 |
| | Age of Acquisition (CW) | 6.27 | 0.61 | .56 |
| | Lexical density (type) | 0.66 | 0.09 | .55 |
| Syntactic | Complex nominals per clause | 1.06 | 0.32 | .49 |
| | Mean length of clause | 9.24 | 1.86 | .46 |
| | Complex nominals per T-unit | 1.97 | 0.91 | .39 |
| | Mean length of T-unit | 17.07 | 6.55 | .33 |
| Discourse | Coordinating conjuncts | 0.01 | 0.01 | -.51 |
| | Basic connectives | 0.05 | 0.01 | -.47 |
| | All causal connectives | 0.02 | 0.01 | -.43 |
| | All demonstratives | 0.03 | 0.01 | -.38 |

*Note.* All correlations are significant at the .01 level (2-tailed)

Table 6.4: Highly Correlated Verbal Complexity Indices with the Participants' Ratings of GA Videotext Difficulty

| Category | Feature | *M* | *SD* | *r* |
|---|---|---|---|---|
| Acoustic and phonological | Pitch (SD) | 67.24 | 27.18 | -.27 |
| | Phonation time | 110.01 | 65.03 | .22 |
| | Frequency of syllables | 479.02 | 286.24 | .22 |
| Lexical sophistication and diversity | MTLD MA wrap (CW) | 74.83 | 35.03 | .34 |
| | Root TTR (AW) | 7.51 | 1.45 | .33 |
| | AWL Sublist-3 (Normed) | 0.004 | 0.006 | .25 |
| | Age of Acquisition (CW) | 6.09 | 0.59 | .23 |
| Syntactic | Complex nominals per clause | 1.03 | 0.41 | .28 |
| | Mean dependency distance | 26.11 | 44.51 | .22 |
| | Mean length of clause | 9.54 | 2.63 | .22 |
| Discourse | All additive (normalized) | 0.05 | 0.02 | -.51 |

*Note.* All correlations are significant at the .01 level (2-tailed)

To explore the contribution of different complexity feature sets to videotext difficulty, four regression models based on selected features from each verbal complexity category were developed. The best model was the lexical regression model ($RMSE = 2.85$) and it accounted for 37% of variance in the language learners' ratings of videotext difficulty. Next best model was the phonological regression model $R^2 = .26, RMSE = 3.07$), followed by

Table 6.5: Summary for Regression Model for Predicting Videotext Difficulty

| Features | $\beta$ | SE | $t$ | Sig. |
|---|---|---|---|---|
| Pitch (SD) | -0.55 | 0.14 | -3.92 | <.001 |
| K1-4 (CW) | 0.57 | 0.15 | 3.70 | <.001 |
| AWL Sublist-3 | 0.47 | 0.15 | 3.05 | .022 |
| AFL | -0.71 | 0.15 | -4.86 | <.001 |
| ASWL Sublist-2 | -0.43 | 0.17 | -2.61 | <.001 |
| Root RTT (FW) | 0.64 | 0.15 | 4.21 | <.001 |
| Complex nominals per clause | 0.42 | 0.17 | 2.49 | <.001 |
| Age of Acquisition (CW) | 1.15 | 0.19 | 5.94 | <.001 |
| COCA spoken bigram AC | 0.47 | 0.15 | 3.23 | <.001 |

*Note.* Constant term= 16.97; $\beta$ is the standardized beta coefficient; SE is standard error

discourse regression model ($R^2 = .17, RMSE = 3.28$). The model with the least explaining was the syntax model ($R^2 = .16, RMSE = 3.30$).

### 6.3.4 Performance of Regression Models

To investigate how the verbal complexity features from the regression analysis on the training set perform on the held-out test set (n=128 videotexts), the regression beta coefficients and constant term from the training were used to predict videotext difficulty in the test set. The findings demonstrated that the regression model yielded a correlation of .59 and an *RMSE* of 2.75 and that the selected verbal complexity collectively explained 33% of the variance in language learners' judgments of difficulty in the testing set. I also evaluated the regression models trained on the four verbal complexity feature sets that were used to predict videotext difficulty in the testing set. As presented in Table 6.6, the model with all features set outperformed the models in predicting the participants' rating of difficulty in videotexts.

Table 6.6: Summary of Regression Models Performance in 10-fold Cross Validation

| Model | $r$ | $R^2$ | RMSE |
|---|---|---|---|
| Acoustic-based model | .52 | .27 | 2.88 |
| Lexical-based model | .56 | .30 | **2.81** |
| Syntactic-based model | .34 | .08 | 3.14 |
| Discourse-based model | .36 | .09 | 3.13 |
| All features model | .59 | .33 | **2.75** |

### 6.3.5 Genre Specific Regression Models

To explore the effect of video genre in videotext difficulty prediction, genre-based regression models were developed. To remove insignificant features, a cross-validated lasso model on AL and GA subcorpora selected 29 and 16 potential verbal complexity features for AL and GA subcorpora, respectively. The significance of these features and whether they have multicollinearity were checked. Any indices that were highly collinear ($r \geq .9$) were identified and the index with the strongest correlation with videotext difficulty scores for a particular corpus was retained, and the other indices were disregarded. After removing insignificant features, the remaining features were included as predictor variables in multiple regression models. The AL regression model yielded a significant model, $F(11, 202) = 26.48, p < .001, r = .79, R^2 = .59$ for the training set. As shown in Table 6.7, the model includes 9 lexical complexity measures: a single measure of acoustic complexity, and a single measure of syntactic complexity. The features collectively explained 59% of the variance in the participants' ratings of difficulty in academic lecture videotexts.

Table 6.7: Summary for Regression Model for AL Videotexts

| Features | $\beta$ | SE | $t$ | Sig. |
|---|---|---|---|---|
| Pitch (SD) | -0.48 | 0.196 | -2.457 | .015 |
| k1-3 (CW) | 0.783 | 0.208 | 3.764 | <.001 |
| AWL-Sublist 4 | 0.417 | 0.191 | 2.184 | .031 |
| AFL | -0.692 | 0.185 | -3.74 | <.001 |
| Log TTR (AW) | 0.565 | 0.208 | 2.714 | .007 |
| HDD (FW) | 0.446 | 0.203 | 2.2 | .029 |
| T-units per sentence | -0.536 | 0.176 | -3.04 | .003 |
| Age of Acquisition (CW) | 1.031 | 0.231 | 4.462 | <.001 |
| COCA spoken bigram AC | 0.826 | 0.209 | 3.956 | <.001 |
| COCA spoken trigram MI2 | 0.436 | 0.211 | 2.066 | .044 |
| COCA academic trigram 2 T | -0.507 | 0.194 | -2.611 | .011 |

*Note.* Constant term = 16.8411; $\beta$ is the standardized beta coefficient; SE is standard error

The regression model for GA subcorpus yielded a significant model, $F(8, 205) = 10.26, p < .001, r = .53, R^2 = .29$, and it included 7 lexical complexity features and a single measure of syntactic complexity. The model coefficients are presented in Table 6.8:

Table 6.8: Summary for Regression Model for GA Videotexts

| Features | $\beta$ | SE | $t$ | Sig. |
|---|---|---|---|---|
| K2 BNC-COCA | -0.83 | 0.23 | -3.64 | <.001 |
| AWL Sublist 3 | 0.70 | 0.22 | 3.08 | .002 |
| AWL Sublist10 (Normed) | 0.43 | 0.20 | 2.12 | .036 |
| Root TTR (AW) | 0.86 | 0.21 | 3.94 | <.001 |
| MTLD (FW) | 0.72 | 0.21 | 3.41 | <.001 |
| Complex nominal per clause | 0.56 | .238 | 2.36 | .019 |
| Meaningfulness (AW) | -0.62 | 0.21 | -2.92 | .004 |
| COCA Spoken trigram MI | -0.66 | 0.20 | -3.23 | <.001 |

*Note.* Constant term= 16.626; $\beta$ is the standardized beta coefficient; SE is standard error

The performances of the AL and GA regression models were evaluated on unseen hold-out testing set for each video genre. Specifically, the constant and B weights of the two trained models were used to predict the difficulty scores of the videotexts in the testing sets. The results indicated that the AL model yielded an *RMSE* of 2.47 and verbal complexity features used in the model collectively explained 50% ($r = .73, R^2 = .50$). The GA model achieved an *RMSE* of 2.83 and its features explained 25% ($r = .52, R^2 = .25$). The performances of the two genre-specific regression models are compared with baseline models developed using the Flesch-Kincaid reading ease formula in Table 6.9.

Table 6.9: A Comparison between the Performances of the two Regression Models and Baseline models on Predicting Videotext Difficulty in the Testing Sets

| **Models** | **r** | $R^2$ | **RMSE** |
|---|---|---|---|
| Baseline model (AL) | .32 | .10 | 3.38 |
| Baseline model (GA) | .04 | -.02 | 3.41 |
| AL Model | .73 | **.50** | **2.47** |
| GA Model | .52 | **.25** | **2.83** |

As shown in Table 6.9, the AL and GA regression models outperformed their baseline models by a significant margin, 33.5% and 19%, respectively.

## 6.4 Discussion

The current study examined the relationship between verbal complexity and language learners' subjective ratings of videotext difficulty, and whether verbal complexity can accurately predict videotext difficulty. The study also investigated the effect of video genre on videotext difficulty prediction. To this end, a set of computational linguistic tools and techniques were employed to extract a variety of verbal complexity features from the audio streams and phonotactic transcripts of the videos in the SLVC corpus. The extracted features account for different dimensions of verbal complexity, including acoustic and phonological complexity, lexical sophistication and diversity, syntactic and grammatical complexity, and discourse cohesion.

The results, overall, suggest that verbal complexity plays an important role in gauging difficulty in L2 videotexts. Specifically, the results demonstrated that verbal complexity explained 37% of the variance in learners' ratings of videotext difficulty for both video genres (AL and GA). The results also showed what verbal complexity features contribute to videotext difficulty varies between the two video genres. A regression model trained on AL videotext yielded a significant model explaining 51% of the variance in learners' ratings of difficulty in this video genre, whereas a model trained on GA videotexts explained about 25% of the variance in learners' ratings of difficulty in GA videotexts. The finding lends support to the notion that video genre has an impact on videotext difficulty assessment—that is, videotext difficulty is better modeled using verbal complexity features related to each video genre. In what follows, I discuss the results of the study with regard to the research questions.

### 6.4.1 The Relationship between Verbal Complexity and Videotext Difficulty

The first question of the study asked whether there is a relationship between verbal complexity and videotext difficulty. The results demonstrated that of the 12 acoustical and phonological complexity indices investigated in the study, only the variance in pitch has proved to be a significant predictor of videotext difficulty. Particularly, videotext difficulty tends to decrease if speakers have considerable pitch variability (pooled mean= 154.21, *SD*= 28.73). This can be explained by the fact that pitch variation impacts attention allocation (Potter, Jamison-Koenig, Lynch, & Sites, 2016) and makes speech more enjoyable,

understandable, and helps listeners to recall information (Hahn, 2004). Variation in pitch work as intonational signals which help the listener to form a coherent mental map of the talk (Thompson, 2003).

As regards the speed of delivery, the results demonstrated no impact for any of the three-speed indices employed in this study (speech rate, articulation rate, and frequency of pausing). This finding is in line with previous studies (Brunfaut & Révész, 2015; Révész & Brunfaut, 2013), who also found no significant relationship between these three measures of speed delivery and spoken text difficulty. The lack of impact for delivery of speed cannot be attributed to low variation because there was a considerable variation in speech rate in our videotexts. The mean speech rate for the 640 videotexts was 3.77 syllables per second (SD= .56, Min= 1.92, Max= 5.35).

This finding suggests that speed of delivery per se may not have a detrimental effect on videotext difficulty. Recent research suggested that speech rate and the amount of information content are inversely correlated; that is speakers tend to produce less informative words and syntactic structures when speaking fast (Cohen Priva, 2017). Another possible explanation is that the effect of speech rate on videotext difficulty may become less significant because of the presence of the speakers. However, further research is warranted to investigate this assumption.

Also, it is worth noting that speaking rate was conceived as a global measure in this study, reflecting the average of speaker's speech rate over the entire videotext, but speech rate is dynamic—that is a speaker may speed up or slow down through a videotext. Therefore, future research should consider using local measures of speed to capture dynamic changes in speed and their impact on videotext difficulty.

The results for phonotactic probability and phonological neighborhood density also showed no significant effect on videotext difficulty. The lack of impact for these two indices on spoken text difficulty has been also reported by previous studies (Révész & Brunfaut, 2013).

In line with previous findings that showed lexical complexity has a significant effect on spoken text difficulty, our results also demonstrated that lexical complexity was a strong predictor of videotext difficulty. Particularly, videotexts with a higher proportion of content words from the BNC-COCA K1-4-word family band proved more difficult. For adequate listening comprehension, language listeners need to know between 2000 and 3000 word families (Van Zeeland & Schmitt, 2013). Furthermore, academic words from the AWL (Sublist-3). On the contrary, videotexts with a high incidence of academic spoken

words (ASWL-Sublist-2) tend to be less difficult for language learners. Further, formulaic sequences in videotexts seem to make videotexts less difficult for language learners.

The remaining lexical complexity measures address different aspects of lexis. A single index of lexical density, root TTR (FW), had a positive relationship with videotext difficulty. This finding suggests that videotext with a greater variety of lexis (here function words) are more likely to impose more cognitive demand on the viewers because they require the recognition of a larger quantity of unique words. These findings for lexical diversity are also consistent with those of Révész and Brunfaut (2013) and Rupp et al. (2001), which also revealed a significant link between lexical diversity and the difficulty of L2 listening. Finally, the mean of AoA of content word norm (Kuperman et al., 2012) was found to be a significant predictor of videotext difficulty in the two video genres.

As far as the complexity of the videotext syntactic structure is concerned, our results showed that a higher ratio of complex nominals per clause to the overall number of clauses in a videotext is a significant predictor of videotext difficulty. Previous studies have generally failed to observe any significant effect of structural complexity on L2 listening difficulty (Blau, 1990; Kostin, 2004; Révész & Brunfaut, 2013). This can be partly ascribed to the small-sized datasets used in previous studies which may not help in revealing useful patterns in the text syntax. The fact that the difficulty of videotext tends to increase as a function of considerable use of complex nominals in the videotext is highly anticipated. Complex nominals are syntactic constructions that include a noun plus an adjective, possessive, prepositional phrase, adjective clause, participles, or appositive; nominal clauses; and gerunds and infinitives in subject position (Cooper, 1976; Lu, 2011a), and because they come before the main verb they potentially place heavier demand on working memory. There is a strong relationship between the number of words before the main verb—for example, garden path sentences—and sentence processing difficulty (e.g., Cunnings, 2017).

Lastly, with regards to the relationship between discourse-level complexity and videotext difficulty, our findings revealed no significant effect of any of the discourse cohesion indices on videotext difficulty. The fact that there is only a handful of relevant studies and there are differences in the types of connectives and how they were investigated make comparing our findings with those of others difficult. However, it is worth noting that our finding is in harmony with the results of Nissan et al. (1995), and with those of Révész and Brunfaut (2013), who observed no association between discourse connectives and perceived difficulty in L2 listening.

While the findings suggest that some verbal complexity features (e.g., phonological complexity and discourse cohesion features) have no direct impact on videotext difficulty, it is, however, very reasonable to assume that their effect on videotext difficulty may have been overshadowed by the presence of many powerful variables in the regression model. Therefore, we need to ask how phonological, lexical, syntactical, and discoursal complexity independently contribute to videotext difficulty?

### 6.4.2 Contributions of Acoustic, Phonological, Lexical, Syntactic, and Discoursal Complexity to Videotext Difficulty

To further examine how each aspect of verbal complexity contributes to videotext difficulty, four regression models with the most predictive features from each category were developed and their performances in accurately predicting videotext difficulty scores in the testing set were compared. The comparison between the four models demonstrated that a model involving only lexical sophistication and diversity indices outperformed the other models in predicting the difficulty in the two video genres. The next best performing model is based on acoustical and phonological complexity features (speech duration, speech rate, articulation rate, and standard deviation of pitch), followed by a model based on discourse complexity features (basic connectives, positive causal, all connective, adjacent overlap FW sentence, adjacent overlap argument sent, and the number of topics), and a model based on a single syntactic complexity index (complex nominals per clause).

The findings, overall, are in accord with our expectation that greater lexical complexity would be associated with increased videotext difficulty. Although the four models had similar performances, we may, however, make some tentative notes on the findings. Lexical sophistication and diversity as well as acoustical and phonological complexity appeared to be a key indicator of videotext difficulty. An explanation for this finding is rather straightforward. Sufficient vocabulary knowledge is very critical for understanding spoken language but recognizing a word in speech depends on how clearly it was articulated. The findings also suggest that syntactic and discourse complexity may not contribute much to videotext difficulty. Reflecting on previous research findings, our findings are consistent with other studies that found low or no significant contribution of syntax and discourse-level features to text difficulty (e.g., Révész & Brunfaut, 2013).

Furthermore, by reflecting upon and comparing the effects of the variables included in the four regression models, we can determine how each predictor is contributing relative

to other predictors. The relative impact of these variables may have gone undetected in our analysis of all features.

### 6.4.3 The Generalizability of Verbal Complexity Models

The generalizability of our verbal complexity features to predict difficulty in the two video genres was evaluated using a 10-fold cross-validation approach. The best model yielded an RMSE of 2.75 and explained 59% of the variance in participants' judgments of videotext difficulty. The model includes acoustic, lexical, and syntactic complexity features and it outperformed models based on predictive features from each verbal complexity category (acoustic and phonological, lexical, syntactic, and discoursal complexity). The two next best predictive models were a model based on lexical complexity features and a model based on prosodic complexity features. When compared to a baseline model, all models did better in predicting videotext difficulty, but the best model (All-feature model) outperformed the baseline model by a significant margin.

Taken together, these results suggest that our genre-agnostic regression model can be used to predict difficulty in academic lectures and government advertisement videotexts with reasonable accuracy. Further, our results demonstrated that simple traditional complexity features did not capture complexity in videotext, and videotext difficulty was better modeled using a variety of verbal complexity features.

### 6.4.4 Genre Specific Regression Models

Our findings lend further support to the notion that verbal complexity varies across text genres (Sheehan et al., 2013, 2014). Specifically, the two regression models trained on the AL and GA subcorpora of the SLVC corpus required different complexity features to accurately predict the difficulty in the two types of videotexts. For example, lexical sophistication and diversity are found to be strong indicators of difficulty in AL videotexts but not a significant indicator of difficulty in GA videotexts. Whereas acoustical and syntactic complexity, as measured by the variation in pitch and the presence of complex nominals per clause, are key indicators of difficulty in GA videotext but not AL videotexts. And when evaluated on a hold-out testing set, the performance of the best-trained models varied considerably (*RMSE*= 2.47 and 2.83, for AL and GA based models, respectively). This finding suggests that the selected verbal complexity features seem to do a better job in explaining and predicting the difficulty in AL videotext but less so in GA videotext.

Unlike instructional videos, government advertisement has not been explored fully in text complexity research; therefore more fine-tuned measures of complexity may be needed to better explain and predict the difficulty in this genre videotext.

Our research was innovative in several aspects. Firstly, we explored the relationship between verbal complexity and videotext difficulty which has not been explored before. Secondly, we interrogated the contribution of a wider range of verbal complexity indices that tap into different dimensions of acoustical and phonological, lexical, syntactical, and discoursal complexity in videotext, of which multiple indices explained a considerable amount of difficulty in L2 videotext. Thirdly, our analysis is based on a large balanced corpus of videotext which was purposefully built for modeling L2 difficulty in two video genres. And lastly, we explored the impact of genre bias on developing automated videotext complexity systems.

Nonetheless, our study has several limitations that need to be acknowledged and addressed in future research. Although we explored different machine learning algorithms to the task of predicting videotext difficulty (e.g., linear SVM, ridge regressor, and elastic net regressor), the performance of our model might have been improved by using non-linear algorithms or by implementing more advanced techniques such as boosting and bagging. Another limitation is that some aspects of verbal complexity were comparatively underrepresented. For example, lexical sophistication and density were represented in the study by much fewer indices than other aspects of lexical complexity.

## 6.5 Chapter summary

In this chapter, I explored the contribution of verbal complexity to videotext difficulty. The findings, overall, demonstrated that verbal complexity can explain and predict videotext difficulty in instructional and government advertisement videotexts. Moreover, I explored whether verbal complexity features can predict language learners' ratings of videotext difficulty. Several indices related to prosodic, phonetic complexity, lexical sophistication, syntactic, grammatical complexity, and text cohesion were extracted from video transcripts and used in regression models. The findings of this study demonstrated that verbal complexity partly predict videotext difficulty and that the difficulty of academic lectures and government advertisement videotexts are better predicted using different sets of verbal complexity features. In the next chapter, I will explore the relationship between visual complexity and videotext difficulty.

# Chapter 7

# The Contributions of Visual Complexity to Videotext Difficulty

In the previous chapter 6, I investigated the relative contribution of verbal complexity to the perception of videotext difficulty. However, videos are dynamic multimodal texts, and it is, therefore, a fair question to ask how visual and spatiotemporal complexity affect videotext understanding. To date, we know little about visual complexity and whether it can explain videotext difficulty in a second language. To address this gap, I investigated the impact of a broadening array of visual and spatiotemporal complexity features, both existing and novel ones, on predicting videotext difficulty as experienced by language learners. Given the explanatory nature of the study, the proposed indices were first subjected to factor analysis, using Principal Component Analysis (PCA), in order to examine whether those features cluster to form latent constructs. Also, in this chapter, I introduce the Automated Video Analysis tool (AUVANA), an open-sourced tool for analyzing, computing, and visualizing visual complexity in videos. In its computations, AUVANA employs a suite of state-of-the-art image processing, computer vision algorithms, and pre-trained AI models to automatically compute several visual complexity indices for modeling and assessing video text difficulty. To validate the proposed indices, they were used for predicting the difficulty of videotexts in the SLVC corpus (AL, GA, and both). Finally, the performances of the developed visual regression models are compared to those of the verbal models developed in Study 1.

Structurally, the chapter is organized into three sections. In the first section, I propose a set of visual complexity features and describe their underlying assumptions, and how they

were extracted and computed. Next, I introduce AUVANA and outline its main functions and utilities. In the third section, I describe the main study of this chapter and report its findings. The chapter ends with a discussion of the results and their implications.

## 7.1 Introduction

When working with video data, it is helpful, if not necessary, to define our units of analysis. Commonly, a video is broken down into its basic temporal units; scenes, shots, and frames. But video researchers have used various units of analysis (e.g., individual utterances, gestures, facial expressions) depending on their theoretical commitments and the specific research questions they are pursuing. In this study, the basic elements of scene, shot, keyframe, and frame were used. As illustrated in Figure 7.1, a single scene can be composed of one or multiple thematically related shots. A single shot is composed of a continuous sequence of frames taken from a single run of a camera, depicting an action over time.


FIGURE 7.1: Video Hierarchical Structure

On a more granular level, we can also talk about keyframes and frames. A keyframe is a single stable frame that best represents the visual content in a shot, whereas a frame is a single still image. In this study, I extracted and computed numerous complexity features from the entire video, a shot, keyframes, a time window of 5 seconds (5sW), and a random sample (RS) of 30 seconds, as shown in Table 7.1.

While it is possible to compute visual complexity features on each frame in the video, it is nonetheless computationally expensive. A more reasonable alternative is the computation of visual features from keyframes especially for features that do not require a focus on temporality. It is also beneficial to ascertain if keyframe-based features can be a less computationally expensive alternative for video-based features. Because the original videos on the SLVC corpus vary in their length, resolutions, and frame per second (FPS) rate, it was important to ensure that all videos are comparable before computing visual

TABLE 7.1: Features Levels

| Features | Video | Shot | 5sW | RS | KF |
|---|---|---|---|---|---|
| Compression | ✓ | | | ✓ | ✓ |
| Colorfuness | ✓ | | ✓ | | ✓ |
| Structural Similarity | ✓ | | ✓ | | |
| Saliency | ✓ | | | | ✓ |
| Clutter | | | | | ✓ |
| Faces | ✓ | | | | |
| Movement | ✓ | | | ✓ | |
| Visual text | ✓ | ✓ | | | |
| Edges | ✓ | | | ✓ | ✓ |
| Objects | | | | | ✓ |
| Production Style | ✓ | | | | |

Note: *5sW* stands for 5 seconds window; *RS* stands for random sample; *KF* stands for keyframe

complexity features. This was achieved by scaling down all videos and keyframes to the same resolution. This step is very important because frame size and video length can directly affect the calculation of several indices of visual complexity (e.g., image size and video compression). It also has helped in increasing the processing and computation of visual complexity features.

In the next section, I will first describe how video shots and keyframes were extracted. Then, I present the visual complexity features used in this study, their underlying assumptions, and how they were extracted and computed.

### 7.1.1 Detecting Shots and Keyframes from Videos

The development of video shot boundary detection (SBD) algorithms has a long history in video processing. SBD algorithms have several applications in areas such as content-based video analysis and retrieval (see Abdulhussain et al., 2018; Satapathy, Govardhan, Raju, & Mandal, 2015, for comprehensive reviews). Formally, the SBD problem can be described as follows: given two consecutive frames, do they belong to the same shot or not. Detecting shot boundaries is a challenging task due to a variety of factors, including video types, spatial effects, motion, and zooming. A transition between shots is typically categorized into two types: abrupt (fast; as hard cut) and gradual (slow; as dissolve, fade-in, fadeout, or wipe) transition (Satapathy et al., 2015).

In the field of video processing, several methods have been proposed to detect shot boundaries in videos containing both types of transitions (Abdulhussain et al., 2018). These approaches, by and large, determine shot transitions if there is a sharp change in color intensity exceeding a certain threshold. For detecting shot boundaries in a videotext, I initially experimented with two SBD algorithms on a small set of videotexts, namely the threshold-based algorithm and the content-aware algorithm, both implemented in the PySceneDetect library. The threshold-based shot detector is by comparing the intensity or brightness of the pixel values in the current frame with a predefined threshold and triggering a shot cut when this value crosses the threshold. The content-aware detector, on the other hand, relies on changes in HSV (hue, saturation, value) color space.

However, a drawback of both approaches—and most SBD algorithms in general—is that they require manual tuning for their threshold parameters. Using different threshold values, I tested the two types of algorithms on a small set of videotexts which remarkably vary in their visual contents. I found the content-aware algorithm yielded the most reliable results with a threshold of 0.3. Therefore, I decided to use the content-aware algorithm for detecting shot boundaries in all videos in the SLVC corpus.

A keyframe extraction refers to the process of selecting a single frame that represents the shot. It can be taken from anywhere in the shot. In this study, I selected the first stable frame from the onset of each shot. In total, 14,297 keyframes were extracted from all the videos in the SLVC corpus.

## 7.2 Computational Measures of Visual Complexity

Image (or frame) complexity generally refers to the notion that "a picture of a few objects, colors, or structures would be less complex than a very colorful picture of many objects that are composed of several components" (Madan, Bayer, Gamer, Lonsdorf, & Sommer, 2018). This working definition is meant for static pictures, but it can be extended to describe spatial or visual complexity in video frames because fundamentally a video is merely a collection of image frames displayed rapidly. Therefore, it would be reasonable to suggest that visual complexity in videos is a function of aggregated complexities within and across video frames. I computed different indices that are pertaining to spatiotemporal complexity in videos, which I will discuss in the following sections.

### 7.2.1 Compression-based Features

In static image complexity literature, researchers have developed various objective metrics to quantify complexity in static pictures. A simple and frequently used quantitative metric for image complexity is the image file size after compression, typically using lossy [1] algorithms such as JPEG and GIF. Generally speaking, lossy algorithms compress data by removing redundant visual information in the picture and optimizing how information is stored and retrieved using more efficient mathematical operations. Therefore, given the same picture dimensions, more visually complex pictures have larger file sizes when compressed than less visually complex pictures. Despite being simple, compression-based measures have been found to correlate with human perception of complexity in still images (Machado et al., 2015; Yu & Winkler, 2013).

Similar to static picture compression algorithms, lossy video compression algorithms aim to reduce or compress the video data so it can be transmitted and shared over the Internet, though how video compression algorithms work is different from those of static pictures due to the temporal nature of video data. The additional dimension of time means that video compression algorithms should eliminate information redundancies in both spatial and temporal dimensions. For example, if a one-minute shot of a person speaking in front of a static background, it would be counterproductive to encode the background image for every frame. A more efficient way is to encode the background frame once and refer back to it until the visual content in the background changes. Besides discarding spatial and temporal redundancies, video compression algorithms rely upon how Human Visual System (HVS) works and capitalize on its lower acuity for color differences (Cohen & Rubenstein, 2020; J.-B. Lee & Kalva, 2008).

There are several types of video compression codecs[2], and while they employ different techniques, they all work toward the same end; reducing file size with as little loss of quality as possible. Video codecs are also designed specifically to meet some practical considerations, for example, online streaming. Practically speaking, there is a trade-off between video quality and the time compression algorithm needs to encode videos, that is a compression algorithm takes a long time to encode video data with high quality. A thorough description of how video compression algorithms work is beyond the scope of this chapter but interested readers can refer to Akramullah (2014).

[1] In contrast to lossy compression algorithms, there are also lossless algorithms that allows for the compression of data without a loss of quality at the expense of large file size.

[2] A technical term which refers to computer hardware or software which enables the compression and decompression of digital data.

To the best of my knowledge, no study has yet examined the use of video compression as an index of Spatio-temporal complexity in L2 videos. Therefore, I used two different video compression codecs: Advanced Video Coding (AVC or H264) and HEVC (High-Efficiency Video Coding or H.265). These two video compression codecs employ different techniques to optimize for video quality and file size (Salomon, 2004; Sayood, 2017), and hence it would be valuable to examine which video compression codec is more indicative of visual complexity in videos.

It is worth noting that because all of the videos in the SLVC corpus were downloaded from *YouTube*, they had already been compressed when uploaded to the site. They also had different lengths, resolutions, bitrate, and frame rates. These factors have an impact on video size when compressed. To ensure comparability between videos in our dataset, three procedures were implemented. First, a 30-second video segment was randomly selected from each video. Second, audio streams were removed, so that they do not weigh on the compressed file size. Third, all videos were scaled to 320 x 240-pixel resolutions, using frame resizing function from the Open Computer Vision Library (OpenCV) with inter-area interpolation. Lastly, I adjusted variable bitrate (VBR) to be between 8-10 MGS and FPS was set to 23 FPS (the average frame rate for all videos was 26 FPS). It is important to note that VBR, as opposed to constant bit rate (CBR), is more dynamic and allocates a higher bit rate for more complex segments in the video and a lower bit rate for less complex segments. This propriety is useful and it allows for accounting for complexity that is due to video content. For compressing the videos, I used the FFmpeg (version: 4.2.2) command-line tool (FFmpeg Developers, 2016).

### 7.2.2 Edge Detection-based Features

Edge detection is a computer vision task that seeks to locate and identify sharp discontinuities in an image, which is typically due to abrupt changes in pixel color intensity (Bhardwaj & Mittal, 2012). These discontinuities in color intensity can help to identify the boundaries of objects in a visual scene. Edge detection is very useful for and often is the first step in many downstream signal processing and computer vision tasks, including, for example, object segmentation, recognition, and image compression.

Because object boundaries are important for human vision, for example, to identify and resolve ambiguity in recognizing objects in a (cluttered) visual scene, it is reasonable to assume that perceived visual complexity would relate to the frequency and distribution of edges in video frames. Measures based on edge detection techniques have been found

to correlate with human perception of image complexity (Machado et al., 2015), but little we know about their potential role in explaining visual complexity in videotexts.

In this study, I used the Canny edge detector, a widely popular edge detection algorithm that is capable of detecting a wide range of edges in an image. It is a multi-stage algorithm proposed by John F. Canny in 1986.

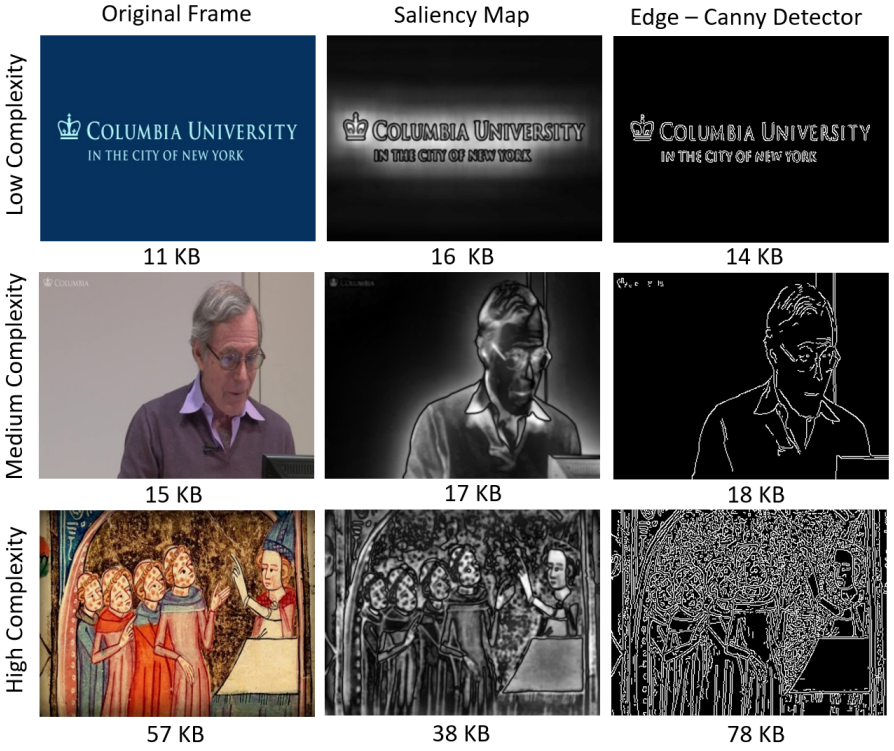



FIGURE 7.2: The Relationship between File Sizes and the Saliency Maps and Edges of Frames of Varying Complexity

I applied the Canny detector on both keyframes and videos. As there is no straightforward way to count edges, I used the file size as a means to quantify edge information in videos and keyframes. To ensure comparability, two procedures were taken; firstly, the video and keyframe were scaled to the same resolution, and secondly, the compression method is kept the same. I computed basic statistics on file size after applying the Canny edge detector on keyframes and videos. Specifically, I summed keyframe file size in KB and calculated the mean, standard deviation, and range, whereas I used the file size in MB to quantify edges in videos.

### 7.2.3 Visual Clutter Features

Several algorithmic methods to quantify clutter in visual displays have been proposed in the computer vision literature (see Moacdieh & Sarter, 2015, for a comprehensive review). In this study, I used two popular metrics developed by Rosenholtz et al. (2007), feature contestation (FC) and subband entropy (SE). The feature congestion algorithm computes how cluttered an image is based on color, luminance contrast, and orientation; whereas subband entropy quantifies the organization of objects in a scene using image encoding efficiency as a proxy. The two algorithms work best on still images. Therefore, I only assessed the visual clutter in the extracted keyframes from each video using a freely available MATLAB script provided by the original authors for both algorithms (Rosenholtz et al., 2007). After applying the two algorithms on each video's keyframes, the values of each metric relating to each video are summed up to yield a single score. I also computed the mean and standard deviation of the clutter values generated by the two clutter algorithms.

### 7.2.4 Colorfulness Features

There is mounting evidence that suggests that color information makes an important contribution to object recognition (Bramão et al., 2011; Tanaka et al., 2001; Witzel & Gegenfurtner, 2018) and invoke emotional reactions. A well-established measure of colorfulness in the image quality literature is the one proposed by Hasler and Suesstrunk (2003). In their study, Hasler and Suesstrunk asked 20 non-expert participants to rate a set of 84 images using a 1-7 scale (ranging from Not Colorful to Extremely Colorful). Then, through running a series of experimental calculations on the collected data, they derived a simple metric that highly correlates with the participants' subjective rating. Specifically, the researchers found a simple opponent color space representation along with the mean and standard deviations of these values correlates to 95.3 % of the participants' ratings. Hasler and Suesstrunk's colorfulness metric has been widely adopted in several research domains and for different applications, for example, in assessing the complexity and anesthetic appeal of websites and visual displays (e.g., Reinecke et al., 2013), and measuring image and video quality (e.g., Winkler, 2012).

In the current study, I computed the colorfulness values for all keyframes, frames extracted every 5 seconds (WS5), as well as on the entire video (frame per second). The

colorfulness algorithm was implemented using Python and OpenCV library. To determine how colorful a frame is, I first split the frame into three colors channels; *Red, Green*, and *Blue*. Then, I use the following set of Equations (7.1) to (7.5) to drive colorfulness metric $C$:

$$re = R - G \tag{7.1}$$

$$yb = \frac{1}{2}(R + G) - B \tag{7.2}$$

$$\sigma_r gyb = \sqrt{\sigma_r g^2 + \sigma_x b^2} \tag{7.3}$$

$$\mu_r gyb = \sqrt{\mu_r g^2 + \mu_y b^2} \tag{7.4}$$

$$C = \alpha_r gyb + 0.3 \times \mu_r gyb \tag{7.5}$$

In equation 7.1, $rg$ is the difference of the Red channel and the Green channel. In equation 7.2, $yb$ represents half of the sum of the Red and Green channels minus the Blue channel. Next, the standard deviation of $rgyb$ and mean of $rgyb$ are computed in equations, 7.3 and 7.4, receptively, before calculating the final colorfulness metric, $C$, in equation 7.5. To derive a global colorfulness metric, $GC$, for the entire video, I simply summed up all colorfulness values computed in the previous step, as shown in equation 7.6:

$$GC = \sum_{i=1}^{n} C_i \tag{7.6}$$

While $GC$ estimates the degree of colorfulness in the entire video, it did not capture abrupt changes in color in two consequent frames. Abrupt changes in color values may impact visual perception. Therefore, I computed the differences in colorfulness between two adjacent frames, ADJGC, using the equation in 7.7:

$$ADJGC = \sum_{i=1}^{n} (C_{t-1} - C_t)^2 \tag{7.7}$$

Additionally, for both $GC$ and $ADJGC$, I calculated the average and standard deviation of $C$ values derived from keyframes, frames extracted every 5 seconds (WS5), and the entire video.

### 7.2.5 Structural Similarity Features

Structural Similarity (SSIM) index is an image quality assessment metric for evaluating the visual similarity between two images (Z. Wang, Bovik, Sheikh, & Simoncelli, 2004). Theoretically, SSIM is used on two similar images, a reference image $x$ and a test image $y$, to quantify their visual similarity. SSIM is computed in several steps, as shown in the Equations (7.8) to (7.11) below:

$$l(x,y) = (2\mu_x\mu_y + C1)/(\mu_x^2 + \mu_y^2 + C1) \tag{7.8}$$

$$c(x,y) = (2\sigma_x\sigma_y + C1)/(\sigma_x^2 + \sigma_y^2 + C2) \tag{7.9}$$

$$r(x,y) = (\sigma_x y + C3)/(\sigma_x\sigma_y + C3) \tag{7.10}$$

$$SSIM(x,y) = [l(x,y)]^{\alpha}.[c(x,y)]^{\beta}.[r(x,y)]^{\gamma} \tag{7.11}$$

where $C1$, $C2$, and $C3$ are constant terms added to avoid instabilities when $(\mu_x^2 + \mu_y^2)$, $(\sigma_x^2 + \sigma_y^2)$, or $\sigma_x\sigma_y$ is close to zero. The $l(x,y)$ index is related with luminance differences, $c(x,y)$ with contrast differences, and $r(x,y)$ with structure variations between $x$ and $y$. The parameters $\alpha, \beta$, and $\gamma$ are added to the general form of SSIM equation in **??** to estimate the relative importance of each component. I used the Scikit-image Python library's implementation of SSIM (version: 0.19.0 van der Walt et al., 2014). Note that the SSIM algorithm takes in two consecutive frames which must have the same dimensions and output a single value indicating to which degree the two frames are similar. A value of 0 indicates the two frames are totally different, whereas a value of 1 indicates the two frames are identical. Then, to derive a single global score for each videotext, I summed up all SSIM values as shown in equation 7.12 :

$$Global\ SSIM = \sum_{i=1}^{n} SSIM_i \tag{7.12}$$

To compute the adjacent score ADJSSI, the equation in 7.13 was used:

$$ADJSSI = \sum_{i=1}^{n} (SSI_{t-1} - SSI_t)^2 \tag{7.13}$$

Like colorfulness features, the SSI values were computed on keyframes, frames extracted every 5 seconds, and on the entire video.

### 7.2.6 Human Faces Features

Both anecdotal and scientific evidence suggests that human faces attract more attention than any other object in the scene. I computed simple indices related to the appearance of human faces in videos, for example, the frequency of unique human faces that appeared in the video. It should be also noted that the recognized faces are most likely the faces of actual speakers in the video, but they can also include any human face displayed in the video. As the appearance of the non-speaking faces may distract the viewer, I calculated the ratio of speaking faces to non-speaking faces. Further, the ratio of appearance/speaking time for the speaker to total running time. If there was more than one speaker, the duration of their appearance time is summed up and averaged to give a single score. For detecting faces, I used OpenCV's pre-trained face recognition classifier based on the Haar cascades algorithm. The cascade classifier has been trained on numerous positive and negative images of face objects and it can subsequently be used to detect faces in other images or videos. One key advantage of this approach is that it is fast compared to deep learning-based classifiers.

### 7.2.7 Object Recognition Features

The quantity of objects shown in video frames is more likely to affect the speed and efficacy of successfully processing and recognizing objects, hence making understanding video more difficult. In the last few years, deep learning-based object recognition models have made great strides in terms of prediction accuracy. Object detection and extraction models take an image or video feed (frame-by-frame) as input and output a predicted label for each object along with the model's accuracy values (confidence) corresponding to each object classified by the model. Object detection and extraction is a very time and memory-consuming process, especially in CPU powered computers. Therefore, in this study, I only considered detecting objects from all keyframes extracted from each video. To this end, I used the ImageAI library's implantation of RetinaNet, a pre-trained deep learning model that was trained on the COCO dataset (Lin et al., 2014). The pre-trained model can detect 80 different everyday objects (e.g., car, chair, cup, and fork). The confidence threshold of the model prediction accuracy was set to 50%, meaning that any object prediction below this confidence threshold was disregarded. To speed up the process, the object detection was run on Google's Tensor Processing Unit (TPU) in Google Collab.

Based on the model output, the quantity of all detected objects as well as unique objects per video and shot were computed. Additionally, the average and standard deviation of all objects and unique objects per video were computed. I also computed the ratio of objects per shot.

### 7.2.8 Motion Detection and Tracking Features

Moving objects attract attention in a visual scene (Wolfe & Horowitz, 2004, 2017). Therefore, I hypothesized that the frequency of moving objects in videos as well as their on-screen duration may distract visual attention, hence contributing to the perception of difficultly in videotext. There are several techniques to detect moving objects in a video stream, including optical flow-based techniques, background subtracting (KaewTraKulPong & Bowden, 2002; Zivkovic & Van Der Heijden, 2006), and frame differencing methods. By and large, motion detection techniques assume that the background of a video stream is largely static and does not change over consecutive frames in videos. Therefore, to detect motion in a video stream, the background should firstly be differentiated from the foreground (see, Sehairi, Chouireb, & Meunier, 2017; Y. Xu, Dong, Zhang, & Xu, 2016, for a review and comparison of different techniques).

In this study, moving objects were detected using the frame differencing method because it is time and memory-efficient. This technique detects motion, $M$, by calculating the absolute difference in pixel intensity between two adjacent frames. That is, in a binary or gray image, for each pixel with coordinates $(x, y)$ in frame $I$, we computed the absolute difference with its corresponding coordinates in the next frame $I_t$ using the following equation 7.14:

$$M(x, y) = |I_t(x, y) - I_{t-1}(x, y)| \tag{7.14}$$

Using pixel coordinates, contours are drawn around each moving object in each frame and then tracked over the subsequent frames until the movement halts. The duration of each moving object is logged in milliseconds. This information allows us to compute some useful indicators of motion in videos, such as the frequency of moving objects, and computed their mean and standard deviation in each video. Finally, to compute a global

measure of motion GM, we use the following formula:

$$GM = \sum_{i=1}^{n} m_i \tag{7.15}$$

where $m$ is the local motion duration in seconds and $d$ is the total duration of the video in seconds.

### 7.2.9 Visual Text Features

The term visual text is used in this study to refer to any letters and numbers displayed in the video stream. A large number of visual texts can be distracting and may lead to information overload. Therefore, I hypothesized that there is an interaction between videotext difficulty and the number of visual texts displayed in the videotext. Computationally, visual texts can be detected and recognized through the use of Optical Character Recognition (OCR) systems. Traditional OCRs are based on carefully crafted features, whereas in modern deep learning-based approaches, informative features are learned from massive training data (e.g., Jaderberg, Vedaldi, & Zisserman, 2014; X. Zhou et al., 2017). Text detection in video streams, as opposed to text documents or still images, is a challenging task due to several factors, including, for example, motion, shadow, lighting, viewing angles, and complex backgrounds which all make detecting and extracting text from its background difficult (see Mancas & Gosseli, 2007, for a comprehensive review).

To detect visual texts, I used the Efficient Accurate Scene Text detector (EAST) proposed by X. Zhou et al. (2017), based on the OpenCV implementation of the detector. EAST is a specialized fully convolutional neural network that has been configured for detecting text from still images and videos. Based on the EAST outputs, basic statistics, e.g., the frequency, average, and standard deviation, of visual texts in the entire video as well as in each individual video shot were computed. Additionally, the ratio of visual texts per shot was considered to be a potential predictor of visual complexity.

### 7.2.10 Saliency Features

When viewing an image without a task in mind, a human typically fixates their eyes on regions or objects that stand out amongst their surroundings. These objects or areas are often referred to as salient and, what makes them so has intrigued many researchers

from different fields, including psychology, neuroscience, computer vision, and robotics. In the last two decades, a large effort has been put toward developing computational models that predict saliency in both static and dynamic images, some of these models were inspired by neural mechanisms (see Kavak et al., 2017; Q. Zhao & Koch, 2013, for a comprehensive review). However, there is no a universally agreed-upon definition of saliency, and existing computational models vary greatly in how they compute saliency. In their computation, saliency models generally make use of either bottom-up (low-level) image features, or top-down (high-level) features, or a combination of both. The majority of existing saliency models are bottom-up driven and mostly rely on a biologically plausible set of features that mimic early visual processing (e.g., Itti, Koch, & Niebur, 1998).

Itti et al. (1998) developed the first computational model of saliency in 1998 and since then researchers have developed numerous models, These models can be classified, according to their mechanisms for estimating visual attention, into several categories. We are interested here in those models that, from a still image or video, produce a visual saliency map. A saliency map is typically an aggregation of several maps that represent (predict) how salient each area of that image is to a human observer (Frintrop, Rome, & Christensen, 2010).

While the main objective of saliency models is to output a single saliency map from an image or video stream, they differ as to what features to include in computing the saliency map (Koch & Ullman, 1985). Computer vision researchers typically make a distinction between fixation prediction and object/region saliency detection (Borji, 2015; Borji & Itti, 2013). Fixation prediction models aim to predict points in an image where humans might look at in a free-viewing task. On the other hand, object saliency models aim to detect and segment salient objects, for example, by drawing a contour around the object. While the objective of the two types of models is different, the two types of models, in practice, often generate similar saliency maps (Borji, 2015). Another distinction is whether the model is developed for predicting static or dynamic saliency. Intuitively, dynamic saliency models should also account for the time dimension in the video input, for example, by tracking moving objects.

In this study, we consider both static and dynamic saliency as potential predictors of videotext difficulty. Static saliency was computed based on keyframes extracted from the videos, whereas dynamic or motion saliency was computed using the entire videotext. Specifically, three different static saliency models were used for estimating saliency in keyframes. The first is a simple, yet influential saliency model developed by Itti et al. (1998). The other two are the Spectral Residual (SR) method (Hou & Zhang, 2007) and

the Fine-Grained (FG) method (Montabone & Soto, 2010). Figure 7.3 shows a comparison between two saliency maps and their file sizes.

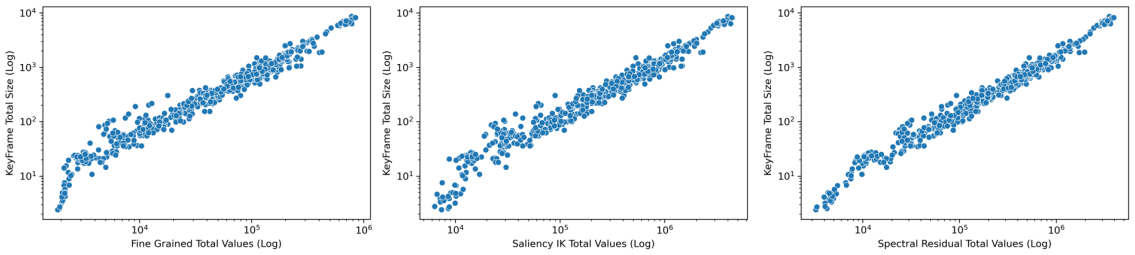



FIGURE 7.3: A Comparison of Saliency Maps Generated by the Three Algorithms and Their Corresponding File Size in KB.

For estimating video saliency, a motion saliency algorithm developed by B. Wang and Dudek (2014) was utilized. This motion saliency detector is based on the background subtraction technique and it considers moving objects as salient objects. Because there is no direct way to quantify saliency, the summation of image and video size of the generated saliency maps for each algorithm was used as a proxy of saliency in keyframes and videotext, respectively. In the case of static saliency, the average, range, and standard deviation were also calculated.

### 7.2.11 Video Production Style

Previous studies in multimedia learning suggest that video instructional styles vary in their pedagogical benefits and some production styles and design principles are more conducive to learning than others (Fiorella & Mayer, 2018; Mayer et al., 2020). It is very plausible, then, to believe that the different production styles in AL videotexts may impact the learner's perception of difficulty differently. To date, there has been no systematic effort to investigate the link between video production style and language learners' perception of videotext difficulty. In the learning science literature, there are several typologies of instructional video production styles (Crook & Schofield, 2017; Hansch et al., 2015). Here, I adopted a taxonomy proposed by Crook and Schofield (2017) for classifying instructional videotext based on their production styles. The taxonomy includes 16 different configurations or formats of lecture styles which are described below

- **A1 Voice over slides:** A sequence of slides is narrated by a hidden voice.

- **A2 Voice over screencast:** A record of a continuous screen recording (as opposed to discrete and static slides) is narrated by a hidden voice.
- **A3 Writing over slides:** Narrated slides include superimposed the narrator's writing. Graphic annotation is added to one or more static images, implicitly by the speaker.
- **A4 Kahn whiteboard:** Narrated whiteboard includes manual acts of superimposed writing. This is similar to A3, except that speaker's hand is made visible as they perform the annotation, thereby conveying a stronger sense of agency.
- **B1 Fixed frame outside:** Video narrator in a window fixed adjacent to a slide sequence. The first of four formats explore picture-in- picture presence of the lecturer. These may each vary in size but are generally small, typically occupying 20% of screen space.
- **B2 Mobile frame outside:** Video narrator in a window in various positions adjacent to the sequence of background presentation activity.
- **B3 Fixed but overlapping:** Video narrator at fixed position but overlapping the background sequence rather than being a framed picture in picture.
- **B4 Mobile frame and overlapping:** Video narrator is now framed but presented at varying positions in the background sequence.
- **C1 Presence in split screen:** Video narrator and slide sequence are presented simultaneously and in adjacent frames.
- **C2 Presence in picture:** Video narrator is visually integrated with slide images as if standing in front of a display surface.
- **C3 Presence overlapped by content:** Symbolic material is superimposed on a video narrator.
- **D1 Presence active on whiteboard:** Narrator moves in front of content and acts upon it, but visual presence overlaps a full-screen presentation surface.
- **D2 Presence in lecture:** Direct recording of narrator in traditional lecture context. The continuity of speaker and display surface is broken, conveying an in-room sense of the two.
- **D3 Presence in full screen:** Close up on a solitary narrator in local 'domestic' or topic-relevant context.

- **E1 Presence in interview:** recorded interview.
- **E2 Presence in discourse:** recorded conversation. This and E1 correspond to more traditional 'talking heads' formats common in broadcast expositions.

Using this taxonomy, I categorized AL videotexts based on their production styles. Of the 16 different formats, there were about 7 different production styles in the AL corpus (see Figure 7.4 for a visual representation of all production styles).


FIGURE 7.4: A Visual Representations of Video Production Styles

Because there is no similar classification for GA videotexts, this feature was not considered in modeling difficulty in the GA video genre.

## 7.3 The Development of AUVANA

To streamline the process of extracting and computing the visual complexity indices described above, I developed the Automated Video Analysis tool (AUVANA). To the best of my knowledge, there is no freely available tool for automatically assessing visual and spatiotemporal complexity in videotext. Therefore, AUVANA was developed to contribute to the scientific understanding of otherwise hard-to-define visual attributes of complex videotexts. Researchers interested in exploring visual and spatial-temporal complexity in videotexts can use the tool to automatically extract, compute, and visualize numerous

visual complexity features. AUVANA is a cross-platform desktop application and it employs state-of-art algorithms and methods for computing a variety of visual complexity features. The tool's interfaces are shown in Figure 7.5, Figure 7.6, Figure7.7 and Figure 7.8

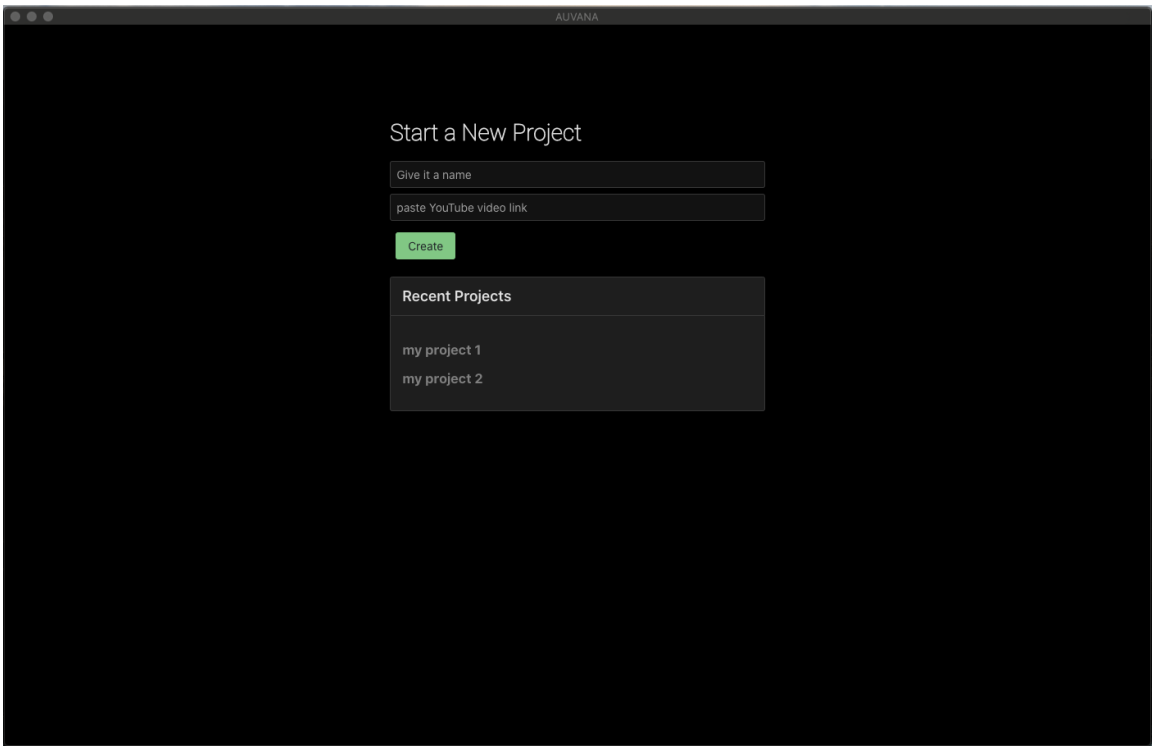



FIGURE 7.5: AUVANA: Main Interface Where Users are Allowed to Upload a Video for Analysis

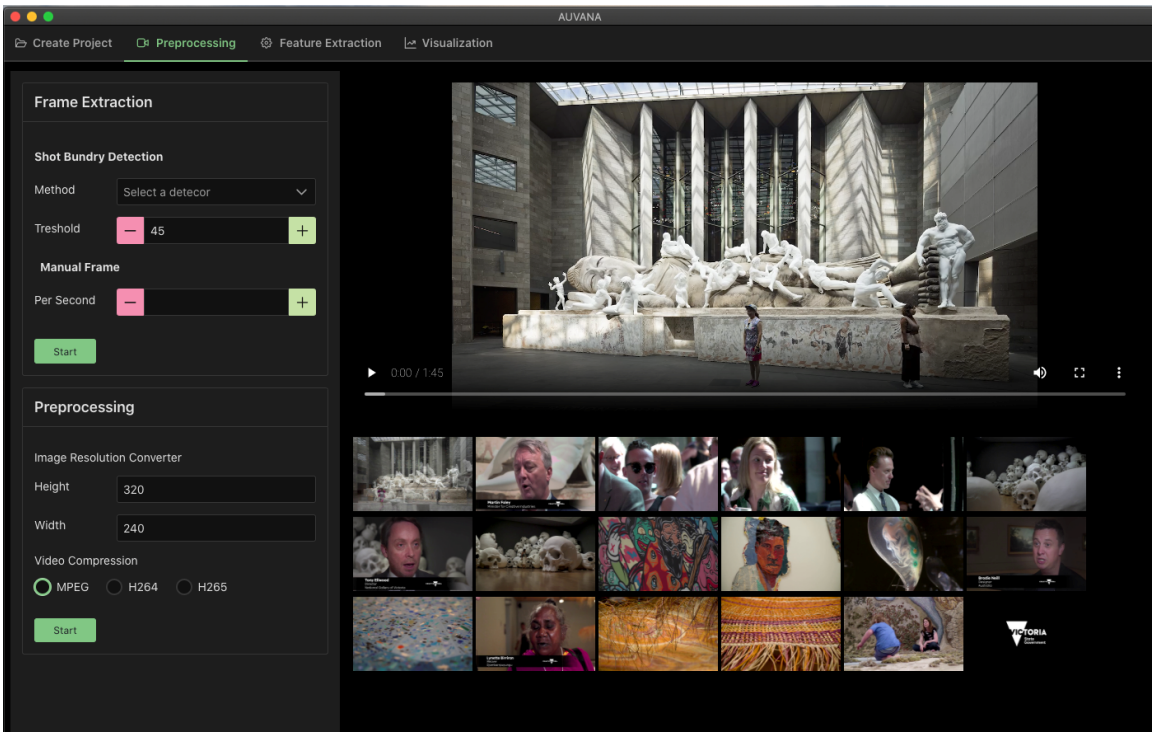



Figure 7.6: AUVANA: Preprocessing and KeyFrame Extraction Interface

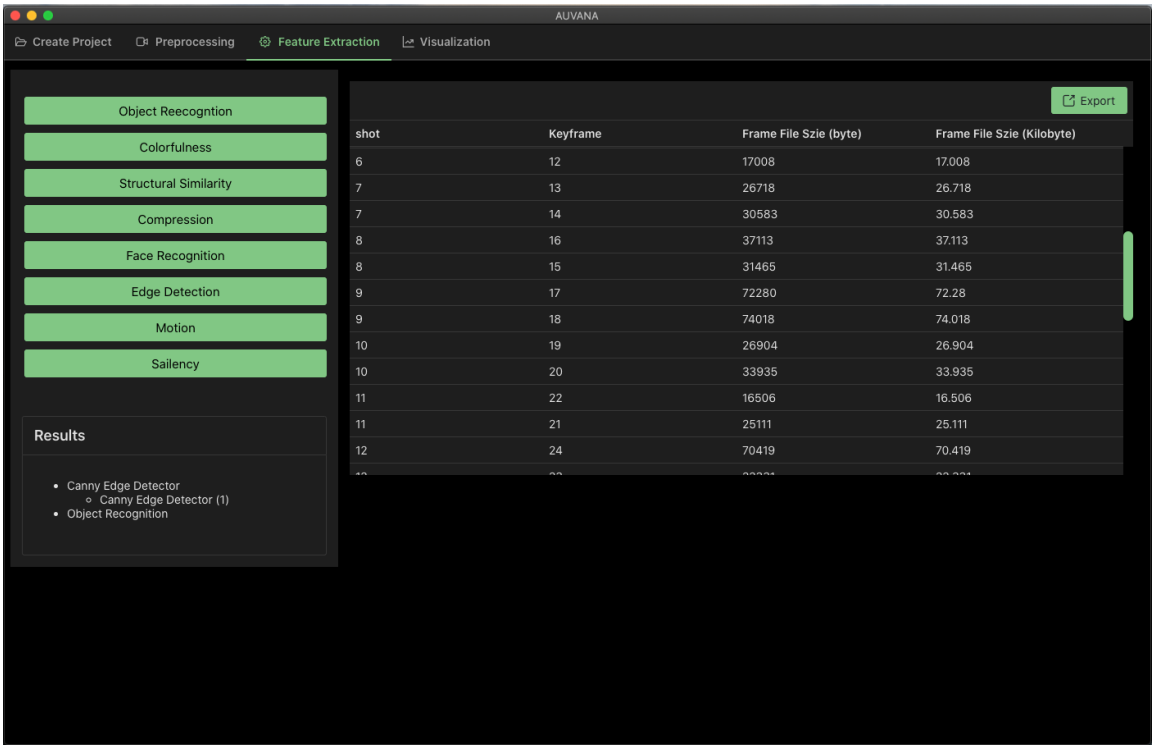



Figure 7.7: AUVANA: Feature Extraction and Computation Interface

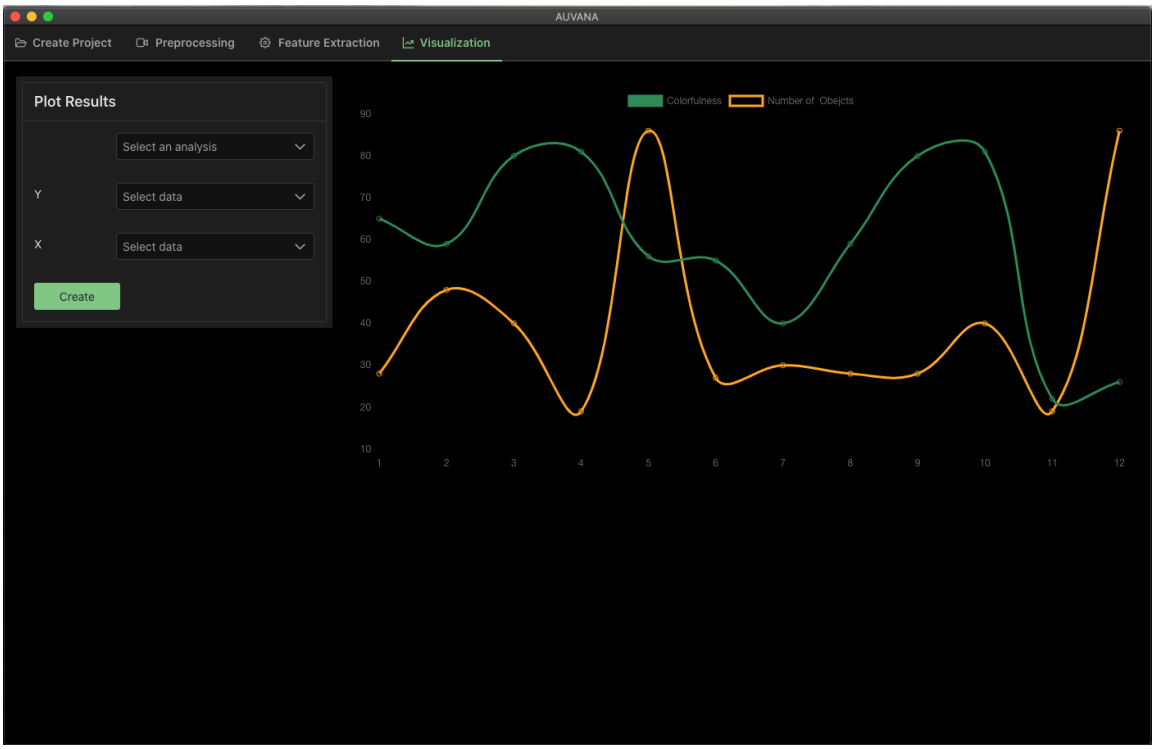



Figure 7.8: AUVANA: Result Visualization Interface

Using AUVANA, I extracted and computed all visual complexity features investigated in this study.

## 7.4 Statistical Procedure and Analysis

### 7.4.1 Research Questions

The current study addresses the following research questions:

1. How do the proposed measures of visual complexity compare with and relate to each other as indices of difficulty in videotext?
2. What is the relationship between visual complexity, and language learners' ratings of videotext difficulty?
3. What visual and spatiotemporal videotext complexity features are more predictive of videotext difficulty?

4. Can visual complexity features accurately predict difficulty in AL and GA videotexts, and both?

### 7.4.2 Corpus

In this study, the entire SLVC corpus was used for developing genre-agnostic regression models, whereas the two subcorpora, academic lecture (AL) and government advertisement (GA), were used for building genre-specific models for each video genre.

### 7.4.3 Exploring Features Dimensionality

Given the exploratory nature of the current study, a large set of potential visual complexity features were considered for explaining and predicting difficulty in videotext. To address the first research question (how visual complexity features compare and relate to each other), the proposed features were analyzed using principal component analysis followed by Direct Oblimin (a non-orthogonal) rotation because it was expected that some of the features are moderately or highly correlated with each other. Before running PCA, features that are derivative of other features were excluded and the remaining features were scaled because PCA is sensitive to the scaling of the original features. Then, different PCA solutions were explored. The most appropriate PCA solution had 11 components as indicated by a scree plot. The sampling adequacy as determined by the Kaiser-Meyer-Olkin test was .83—which is above the commonly recommended value of .6—and Bartlett's test of sphericity was significant ($\chi_2(1770) = 70920.99$, $p < .01$).

### 7.4.4 Data Preprocessing and Feature Selection

After extracting and computing our visual complexity features, the entire dataset was shuffled and divided into a training set and test set following the 67/33 split approach (approximately 67 percent training, 33 percent test). This resulted in a training set of 428 videotext samples and a testing set of 128 videotext samples. The same procedures were applied to the AL and GA subcorpora, which resulted in 214 videotexts for training and 106 videotexts for testing for each video genre. We conducted an explanatory data analysis and data preprocessing for the training sets (e.g., standardization, converting categorical feature to numerical, imputing mi sing values, etc.). The results of the

correlation analyses showed that some features moderately to strongly correlated with the dependent variable. The results can be found in Appendix A.9

Besides using PCA for evaluating the proposed features, PCA is useful for reducing the dimensionality of our dataset to fewer components (or factors) that capture most of the variance in the dataset. To select informative features to be included in the regression models, multiple experiments using PCA, LASSO and, random forest were conducted. The results showed that features selected by the LASSO algorithm are more predictive than the other approaches. The LASSO lambda was determined in a 10-fold cross-validation and the best model selected 28 features for the AL subcorpus, 10 features for the GA subcorpus, and 35 features for both corpora.

### 7.4.5 Models Building and Evaluation

For understanding the relationship between visual complexity and videotext difficulty, simple OLS regression was used. Our goal was to find parsimonious, easy interpretable models that can explain large amounts of variance. After fitting the model in the training set, features with low $p$ values were dropped. Multicollinearity among the features was checked using the Variance Inflation Factor (VIF). The remaining features were then used to predict the difficulty of videotext in the holdout testing sets. Residual plots and Q-Q plots of residuals were visually inspected to review homogeneity of variance (homoscedasticity) and normal distribution of residuals. To develop predictive models, I experiment with different regression algorithms such as OLS regression, LASSO, Ridge, and Support Vector Regression (SVR).

To further investigate the generalizability of the features as well as to avoid sampling bias, the performances of the models were compared using a 10-fold cross-validation method. I consider both adjusted $R^2$ and RMSE as evaluation metrics.

## 7.5 Results

### 7.5.1 Principle Components Analysis (PCA)

The PCA analysis yielded that approximately 85% of the variance in the dataset was accounted for via 11 components. The loadings of the factors and their commonalities are

presented in Table 7.5.1. A conservative threshold of .5 was chosen as a cutoff point. The first component explained 32.28% of the variance and it was comprised of fourteen visual complexity features. Specifically, features in this component estimate the total values of visual clutter, saliency, and compression. The second component summarizes variance detected via four features; the number of speakers, objects, and shots per minute, and the number of speakers in the video. Five features related to visual texts loaded heavily on the third component. The fourth component had nine visual complexity features that estimate the standard deviation and range of visual saliency, clutter, detected edges, and keyframe file sizes. Five measures of colorfulness and structural similarity loaded heavily on the fifth component. The average saliency, clutter, and keyframe file size made the sixth component. The remaining components, 7, 8, 9, 10, and 11 had features that, respectively, estimate the number of detected objects in the video, motion duration, number of speakers, number of faces, and number of moving objects. The correlation matrix showed that the components did not have a higher correlation among each other (see Appendix A.7 for correlation matrix).

| Feature | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | Com. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of Shots | 0.98 | - | - | - | - | - | - | - | - | - | .98 |
| Saliency SR KF Sum | 0.96 | - | - | - | - | - | - | - | - | - | .96 |
| Saliency FG KF Sum | 0.96 | - | - | - | - | - | - | - | - | - | .97 |
| Canny KF KB Sum | 0.94 | - | - | - | - | - | - | - | - | - | .97 |
| Number of Unique Objects KF | 0.94 | - | - | - | - | - | - | - | - | - | .88 |
| Number of Objects KF | 0.87 | - | - | - | - | - | - | - | - | - | .87 |
| Clutter SE Sum | 0.81 | - | - | - | - | - | - | - | - | - | .93 |
| Saliency IK sum | 0.8 | - | - | - | - | - | - | - | - | - | .93 |
| Video Compression H264 Size MB | 0.79 | - | - | - | - | - | - | - | - | - | .93 |
| Clutter FC Sum | 0.77 | - | - | - | - | - | - | - | - | - | .90 |
| Video Compression H265 Size MB | 0.72 | - | - | - | - | - | - | - | - | - | .75 |
| Number of Moving Objects per Video | 0.65 | - | - | - | - | - | - | - | - | - | .72 |
| Canny Video Size MB | 0.65 | - | - | - | - | - | - | - | - | - | .85 |
| Clutter SE SD | - | 0.85 | - | - | - | - | - | - | - | - | .83 |
| Saliency IK SD | - | 0.79 | - | - | - | - | - | - | - | - | .79 |
| SSI Video W5Sec SD | - | 0.76 | - | - | - | - | - | - | - | - | .65 |
| Clutter FC SD | - | 0.68 | - | - | - | - | - | - | - | - | .69 |
| SSI Video W5Sec mean | - | -0.54 | - | - | - | - | - | - | - | - | .78 |
| Number of Visual Text per Min. | - | - | 0.96 | - | - | - | - | - | - | - | .91 |
| Number of Visual Text per Sec. | - | - | 0.96 | - | - | - | - | - | - | - | .91 |
| Number of Visual Text | - | - | 0.88 | - | - | - | - | - | - | - | .89 |

**Table 7.2 continued from previous page**

| Feature | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | Com. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of Visual Text per Shot | - | - | 0.8 | - | - | - | - | - | - | - | .80 |
| Motion Duration Difference Variance normalized | - | - | - | 0.93 | - | - | - | - | - | - | .84 |
| Motion Duration Difference Variance | - | - | - | 0.91 | - | - | - | - | - | - | .85 |
| Motion Duration Difference SD Normalized | - | - | - | 0.87 | - | - | - | - | - | - | .90 |
| Motion Duration Difference Mean Normalized | - | - | - | 0.82 | - | - | - | - | - | - | .86 |
| Motion Duration Difference SD | - | - | - | 0.8 | - | - | - | - | - | - | .88 |
| Motion Duration Difference Mean | - | - | - | 0.72 | - | - | - | - | - | - | .81 |
| Saliency SR KF Mean | - | - | - | - | 0.84 | - | - | - | - | - | .86 |
| Keyframe KB mean | - | - | - | - | 0.83 | - | - | - | - | - | .92 |
| Saliency FG KF Mean | - | - | - | - | 0.82 | - | - | - | - | - | .78 |
| Canny Keyframe KB Mean | - | - | - | - | 0.8 | - | - | - | - | - | .89 |
| Colourfulness KF mean | - | - | - | - | - | -0.89 | - | - | - | - | .87 |
| Colourfulness Adjacent KF mean | - | - | - | - | - | -0.85 | - | - | - | - | .82 |
| Colourfulness Video W5Sec mean | - | - | - | - | - | -0.85 | - | - | - | - | .78 |
| Colourfulness KF SD | - | - | - | - | - | -0.79 | - | - | - | - | .87 |
| Colourfulness Video W5Sec SD | - | - | - | - | - | -0.66 | - | - | - | - | .88 |
| Colourfulness Adjacent Video W5Sec SD | - | - | - | - | - | -0.66 | - | - | - | - | .86 |
| Colourfulness Adjacent KF SD | - | - | - | - | - | -0.62 | - | - | - | - | .75 |
| Keyframe KB SD | - | - | - | - | - | - | 0.93 | - | - | - | .91 |
| Saliency SR KF SD | - | - | - | - | - | - | 0.85 | - | - | - | .81 |

**Table 7.2 continued from previous page**

| Feature | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | Com. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Canny Keyframe KB SD | - | - | - | - | - | - | 0.81 | - | - | - | .86 |
| Saliency FG KF SD | - | - | - | - | - | - | 0.75 | - | - | - | .74 |
| Colourfulness Adjacent Video W1Sec SD | - | - | - | - | - | - | - | 0.96 | - | - | .89 |
| Colourfulness Video W1Sec SD | - | - | - | - | - | - | - | 0.93 | - | - | .84 |
| SSI Video W1Sec SD | - | - | - | - | - | - | - | 0.83 | - | - | .76 |
| SSI Video W1Sec mean | - | - | - | - | - | - | - | -0.77 | - | - | .73 |
| Colourfulness Video W1Sec mean | - | - | - | - | - | - | - | 0.69 | - | - | .55 |
| Number of Faces per Shot | - | - | - | - | - | - | - | - | 0.8 | - | .74 |
| Number of Faces per Min. | - | - | - | - | - | - | - | - | 0.74 | - | .83 |
| Number of Faces | - | - | - | - | - | - | - | - | 0.69 | - | .77 |
| Saliency IK mean | - | - | - | - | - | - | - | - | - | 0.86 | .78 |
| Clutter SE Mean | - | - | - | - | - | - | - | - | - | 0.78 | .87 |

## 7.5.2 Regression Models

### 7.5.2.1 All Video Corpus

Multiple regression analysis was conducted to explore if the selected visual complexity features significantly explain the variance in the participants' ratings of videotext difficulty. After removing insignificant features (features with coefficients exceeding the common alpha level of 0.05), the final model yielded a significant model, $F(8, 419) = 21.92, p < .001, r = .54, R^2 = .30$. The model includes 8 visual complexity features which are *Number of Shots, Colorfulness Video W1Sec (SD), Colorfulness Adjacent Video W1Sec (SD), SSI Video W5Sec (SD), Canny Keyframe KB (Sum), Canny Keyframe KB (Sum), Video Compression H264 Subsample Size (MB), Motion Duration Difference (SD), and Motion Duration Difference (SD-Normalized).* The features jointly explained 30% of the variance in the participants' ratings of videotext difficulty in the training set. Table 7.3 shows additional information regarding this model and the beta weight for each index in the training model.

Table 7.3: Summary for Visual Regression Model for Videotext Difficulty (ALL)

| Features | $\beta$ | SE | $t$ | Sig. |
|---|---|---|---|---|
| Number of Shots | 1.21 | 0.23 | 5.34 | <.001 |
| Colorfulness Video W1Sec (SD) | 1.03 | 0.33 | 3.12 | <.001 |
| Colorfulness Adjacent Video W1Sec (SD) | -0.81 | 0.34 | -2.38 | .021 |
| SSI Video W5Sec (SD) | 0.54 | 0.15 | 3.56 | <.001 |
| Canny Keyframe KB (Sum) | -0.65 | 0.23 | -2.8 | .013 |
| Video Compression H264 Subsample Size (MB) | 1.08 | 0.23 | 4.79 | <.001 |
| Motion Duration Difference (SD) | 0.87 | 0.24 | 3.68 | <.001 |
| Motion Duration Difference (SD-Normalized) | -0.9 | 0.22 | -4.11 | <.001 |

*Note.* Constant term= 16.90; $\beta$ is the standardized beta coefficient; SE is standard error

To investigate how the visual complexity features from the regression analysis on the training set perform on a held-out test set (n=212), the regression beta coefficients and constant term from the training model were used to predict videotext difficulty in the testing set. The findings demonstrated that the regression model yielded an RMSE of 2.79 and that the selected visual complexity features collectively predicted 32% of the variance in the testing set. The Pearson correlation between the true and predicted videotext complexity scores was $.56, p < .001$.

#### 7.5.2.2 AL Video Corpus

A multiple regression model with 7 visual complexity features was fitted to explain the difficulty in AL videotext. The fitted model yielded a significant model, $F(11, 202) = 18.03, p < .001, r = .70, R^2 = .50$. The visual complexity features used in the model were: *Colorfulness KF (SD), Colorfulness Video W1Sec (SD), SSI Video W1Sec (mean), Saliency SR KF (Mean), Saliency FG KF (Mean), Saliency IK (SD), Canny H264 Subsample (MB), Number of Moving Objects per minute, Motion Duration Difference Subsample (Mean), Frequency of Visual Text per Shot and Number of Shots.* Table 7.4 shows additional information related to the AL model and the beta weights for each index in the training model.

Table 7.4: Summary for Visual Regression Model for AL Videotext Difficulty

| Features | $\beta$ | SE | $t$ | Sig. |
|---|---|---|---|---|
| Number of Shots | 1.3 | 0.24 | 5.34 | <.001 |
| Frequency of Visual Text per Shot | 0.53 | 0.21 | 2.49 | .013 |
| Number of Moving Objects per minute | 0.57 | 0.22 | 2.66 | .010 |
| Motion Duration Difference Subsample (Mean) | 0.68 | 0.22 | 3.18 | <.001 |
| Colorfulness KF (SD) | -0.53 | 0.22 | -2.39 | .021 |
| Colorfulness Video W1Sec (SD) | 1.01 | 0.27 | 3.78 | <.001 |
| SSI Video W1Sec (mean) | 1.61 | 0.44 | 3.67 | <.001 |
| Saliency SR KF (Mean) | -0.59 | 0.25 | -2.4 | .025 |
| Saliency FG KF (Mean) | 0.59 | 0.22 | 2.69 | .018 |
| Saliency IK (SD) | 0.85 | 0.28 | 3.11 | <.001 |
| Canny H264 Subsample (MB) | 1.03 | 0.36 | 2.84 | .017 |

*Note.* Constant term= 17.15; $\beta$ is the standardized beta coefficient; SE is standard error

The trained model was then used to predict videotext difficulty in the test set. The results indicated the model yielded an *RMSE* of 2.52 and that the seven visual complexity features collectively predicted 52% of the variance in the testing set. The Pearson correlation between the true and predicted videotext complexity scores was $.73, p < .001$.

#### 7.5.2.3 GA Video Corpus

The best regression model for explaining and predicting GA videotexts included 4 visual complexity features; *Colorfulness Adjacent KF (mean), Colorfulness Adjacent Video W5Sec, Video Compression H264 Size (MB), and Number of Faces per Minute.* The selected features yielded a significant model, $F(4, 209) = 9.36, p < .001, r = .39, R^2 = .15$. When the trained GA regression model was used to predict videotext difficulty in the held-out testing set,

the results showed that the model explained about 23% of the variance in the language learners' ratings of videotext difficulty. The *RMSE* value for the model was 2.72 and the correlation between the true and predicted difficulty scores was $r = .50, p < .001$. Table 7.5 shows additional information related to the GA model and the beta weights for each index in the training model.

Table 7.5: Summary for Visual Regression Model for GA Videotext Difficulty

| Features | $\beta$ | SE | $t$ | Sig. |
|---|---|---|---|---|
| Number of Faces per Minute | -0.67 | 0.23 | -2.96 | <.001 |
| Colorfulness Adjacent KF (mean) | -0.81 | 0.25 | -3.27 | <.001 |
| Colorfulness Adjacent Video W5Sec | 1.04 | 0.25 | 4.13 | <.001 |
| Video Compression H264 Size (MB) | 0.78 | 0.22 | 3.53 | <.001 |

*Note.* Constant term= 16.75; $\beta$ is the standardized beta coefficient; SE is standard error

#### 7.5.2.4 Models Prediction Performance

To further investigate the generalizability of the feature sets and to avoid sampling bias, the selected feature sets for each video corpus were evaluated using a 10-fold cross-validation approach. In addition to the OLS regressor, I also investigated how more advanced machine learning algorithms, such as linear Support Victor Machine (SVR), LASSO, and Ridge regressors, would perform in predicting videotext difficulty in the three datasets. A comparison of prediction performances of all models is shown in Table 7.6.

Table 7.6: A Comparison of the Prediction Performances of the Four Regression Models Using a 10-fold Cross Validation Approach

| Corpus | Model | $r$ | $R^2$ | RMSE |
|---|---|---|---|---|
| AL | OLS | .65 | .45 | 2.71 |
| | LASSO | .65 | .45 | 2.71 |
| | Ridge | .66 | .45 | **2.70** |
| | SVR | .63 | .41 | 2.79 |
| GA | OLS | .35 | .22 | **2.94** |
| | LASSO | .35 | .21 | 2.97 |
| | Ridge | .33 | .22 | 2.96 |
| | SVR | .37 | .20 | 3.00 |
| ALL | OLS | .50 | 29 | **2.90** |
| | LASSO | .51 | .29 | **2.90** |
| | Ridge | .50 | .29 | 2.91 |
| | SVR | .49 | .28 | 2.93 |

All in all, the results showed that all different regression algorithms performed comparably in predicting videotext difficulty, with an exception of the SVR model which performed slightly worse than other models. The best model for predicting difficulty in AL video corpus was ridge regressor ($RMSE$ = 2.70), though both OLS and LASSO models did comparably well ($RMSE$ = 2.71). For predicting the difficulty of GA videotexts, an OLS regressor achieved the best result ($RMSE$ = 2.94). Finally, the comparison between models trained to predict difficulty in both video types shows that both OLS and LASSO models yielded an $RMSE$ of 2.90.

## 7.6 Discussion

The main objective of Study 2 was to examine whether visual complexity can explain the difficulty of AL and GA videotext, as experienced by language learners. To this end, a set of 78 visual complexity features was proposed as potential predictors of visual complexity in videotexts. The proposed features are theoretically motivated, and they were intended to account for different attributes of videotext, including spatial complexity, saliency, clutter, object recognition, and motion. These attributes are of interest because we do not have yet a concrete definition of what visual complexity in a videotext entails.

A second objective of the chapter was to develop the Automated Video Analysis tool (AUVANA) for extracting and computing the proposed visual complexity features and visualizing their outcomes. To validate the newly proposed features, different regression models were developed to explain and predict difficulty in two types of videotexts (AL and GA) as well as in both types combined (ALL).

Overall, the results of the current study suggested that visual complexity contributes to videotext difficulty and in par with the predictive performance of verbal complexity models reported in Chapter 6. Specifically, the results demonstrated that visual complexity explained about 29% of variance in the participants' judgment of videotext difficulty (both video genres, AL and GA, combined). The findings of the current study have important implications for automatic videotext difficulty assessment and feedback systems. I discuss these findings with regard to the study's research questions.

### 7.6.1 The Multidimensionality of Video Visual Complexity

The first question of this study asked how the proposed visual complexity features relate and compare to each other, and whether they make clusters that may characterize some underlying latent structures. To address this question, a principle component analysis was performed on the proposed visual complexity features. The PCA results demonstrated that the 78 features clustered into 11 components which collectively explained 85% of the variance in the dataset. While it is tempting to name or label the extracted components, this step was skipped to avoid reification—assuming that the clusters of variables represent a latent variable while in reality, they do not. But most importantly, this finding suggests that video visual complexity is a multidimensional and multifaceted construct. Obviously, further research is needed for investigating if the patterns found in our dataset exist in other types of videotext.

### 7.6.2 The Relationship between Visual Complexity and Videotext Difficulty

The second question investigated what visual complexity features can explain the difficulty in the two video genres (AL and GA) as well as in both genres combined. The results of OLS regression analyses have demonstrated that seven visual complexity features explained about 50% of the variance in AL corpus, whereas four features explained

about 15% of the difficulty in GA corpus. The results also showed that the difficulty in the two video genres is best explained by a different set of visual complexity features. This can be attributed to the spatiotemporal differences between the two video genres. To investigate this finding further, I performed t-test analyses on the visual complexity features from both genres and the results showed clearly that there were statistical differences between the two video genres in the vast majority of the features. For example, there was a significant statistical difference between the frequency of video shots in AL ($M$= 12.99, $SD$= 6.68) compared to the frequency of shots in GA is ($M$= 4.12, $SD$= 2.23), $t(638)$= 22.54. $p$= .001. Moreover, differences in colorfulness values between the two types of videos were also noted. Accordingly, accounting for video genre differences in visual complexity assessment is critical for more reliable analyses and findings.

With regards to what visual attributes proved to be strong predictors in AL videotext, the results showed that six of the eight features included in the regression model had significant positive relationships with the rating of videotext difficulty. These features were *Colorfulness Video W1Sec (SD), SSI Video W1Sec (mean), Saliency FG KF (Mean), Saliency IK (SD), Canny H264 Subsample (MB), Number of Moving Objects per minute, Motion Duration Difference Subsample (Mean), Frequency of Visual Text per Shot* and *Number of Shots*. The remaining two features, *Colorfulness KF (SD)* and *Saliency SR KF (Mean)*, had negative relationship with videotext difficulty.

Concerning GA videotexts, the results of regression analysis indicated that four features were significant predictors of difficulty in GA videotext. Three of the features had a significant positive relationship with the dependent variable. These features were *Colorfulness Adjacent Video W5Sec, Video Compression H264 Size (MB)* and *Number of Faces per Minute*. The remaining feature, *Colorfulness Adjacent KF (mean)*, had a negative relationship with the ratings of difficulty of GA videotext.

For explaining the difficulty in both video genres, the results demonstrated that eight visual complexity features can collectively explain approximately about 31% of the total variance. As shown in Table 7.3, six of the eight features had positive coefficients: *Number of Shots, Colorfulness Video W1Sec (SD), SSI Video W5Sec (SD), Canny Keyframe KB (Sum), Canny Keyframe KB (Sum), Video Compression H264 Subsample Size (MB)*, and *Motion Duration Difference (SD)*. Whereas two features, *Colorfulness Adjacent Video W1Sec (SD)*, and *Motion Duration Difference (SD-Normalized)*, had negative interaction with videotext difficulty.

### 7.6.3 Predicting Videotext Difficulty Using Visual Complexity Features

The third and last question was concerned with evaluating the predictive performances of the developed regression models (OLS, LASSO, Ridge, and SVR). To ensure the reliability of the results, the prediction performances of the regression models were validated using a 10-fold cross-validation approach. The results showed that a ridge regressor trained on the AL video corpus yielded the best result, *RMSE* of 2.70; whereas the best performing model for GA videotext was the OLS regressor yielding an *RMSE* of 2.94. A comparison of the predictive performances of the models trained on both types of videotexts demonstrated that both OLS and LASSO did better than the other models, achieving both an *RMSE* of 2.90. Generally, the models slightly vary in their performances within each video genre.

To sum up, the proposed video visual complexity indices were able to partly explain and predict videotext difficulty. In fact, the prediction performances of the visual complexity regression models are comparable to those developed in Study 1 (Chapter 6) using only verbal complexity features. Table 7.7 compares the prediction performance of the best verbal and visual complexity model in each video genre (AL and GA) as well as in both genres combined.

Table 7.7: A Comparison of the Prediction Performances of Verbal and Visual Regression Models

| Models | $R^2$ | $r$ | RMSE |
|---|---|---|---|
| AL Verbal | .5 | .73 | 2.47 |
| AL Visual | .45 | .66 | 2.70 |
| GA Verbal | .25 | .52 | 2.83 |
| GA Visual | .22 | .35 | 2.94 |
| ALL Verbal | .33 | .59 | 2.75 |
| ALL Visual | .29 | .50 | 2.90 |

This finding attests to the fact that both verbal and visual complexity play a role in language learner perception of videotext difficulty, hence, both types of complexity should be considered when selecting or designing videotexts for language learning, teaching, and development. The question of whether integrating both verbal and visual complexity lead to better prediction of videotext difficulty will be addressed in the next chapter.

### 7.6.4 Contributions and Limitations

The present study was innovative in several aspects. First, the study proposed several novel visual complexity features for explaining and predicting difficulty in videotexts. The proposed features are theoretically driven and they tap into different dimensions of visual perception and understanding. The PCA analysis has confirmed the multidimensionality nature of the construct of video visual complexity. Empirically, subsets of these features were incorporated into computational models that explained a considerable amount of variance in video difficulty rating, in par with that of verbal complexity models in 6. The visual complexity models also did well in predicting difficulty in unseen videotexts. The results should encourage further research into explicating the relationship between visual complexity and learners' perception of video difficulty and devising more robust computational spatial-temporal complexity measures. Another contribution of this study is the development of AUVANA, an easy-to-use desktop application for extracting, computing, and visualizing complexity features. The tool is geared towards helping applied linguists and researchers in the social sciences in general who are interested in analyzing videotexts but may lack technical knowledge in more advanced computer vision and image processing techniques.

These contributions notwithstanding, the current study has some limitations that could serve as a basis for new lines of research. First, while I leveraged the state-of-the-art methods and techniques for extracting and computing visual complexity features, it should be noted that the best computer vision systems achieve only rudimentary performance compared to humans in some tasks, e.g., object recognition. Second, there are several other alternative approaches and techniques for going about estimating each of the visual complexity dimensions explored in this chapter. For instance, there are different techniques for compressing video data. Therefore, future research should consider comparing and investigating the usefulness of the different techniques for each task. Third, there are also some limitations as to the way some features are computed. For instance, hand gestures were identified as continuous movements of the speaker's hands but no disinclination was made between the different types of gestures. It has been suggested that iconic gestures are more helpful than other types of gestures when used alongside a description of an action or event because they usually reinforce the semantic meaning of accompanying speech (Dargue et al., 2019). Beat gestures, which are not semantically related to the accompanying speech, seem to improve narrative comprehension (Dargue & Sweller, 2020b; Vilà-Giménez, Igualada, & Prieto, 2019). Given the different roles and communicative functions of gestures, there is a need for research to ascertain what type

of gestures is more beneficial for language learners. Similarly, in this study, all moving objects were identified and counted, regardless of whether they were new or not. Future research should perhaps eliminate the recurrent objects. Because they are already recognized, recurrent objects may have a lesser impact on the viewer's attention. Feature research should also explore if these features are universal; that is, they are generalizable to other types of videos and groups of learners.

Finally, it was assumed that video length impacts the computation of several of the visual complexity measures, much like the problem of the text length dependency in calculating lexical complexity measures. To mitigate this problem, I took a single random sample for computing visual complexity features that are more sensitive or dependent on video length (e.g., compression). There are other resampling techniques such as bootstrapping that might provide more stable estimates, as such future studies may consider investigating which approach is more effective for tackling the video length problem.

Clearly, we just scratched the surface in this chapter, and with the rapid development of new computational approaches, there will be always, of course, room for further improvement of visual complexity estimation. For example, visual attention modeling is a very active research field and numerous approaches have been proposed in the last few years. However, it should be noted that the majority of exiting attention models do not account for the multimodal nature of videos, neglecting the influence of both visual and auditory cues in attention guidance. In a recent study, Sidaty et al. (2017) addressed this problem through integrating both auditory and visual cues in their attention model which achieved better modeling of visual attention.

## 7.7 Chapter Summary

In this chapter, I extracted and computed a broadening array of visual complexity features for explaining and predicting difficulty in videotext. The features were extracted using a variety of algorithms and techniques from the fields of computer vision and signal processing. To examine if the features represent underlying latent structures, a principal components analysis was performed on the entire dataset. The results have shown that the proposed features are grouped into 11 different clusters or components, suggesting that the clustered features assess different aspects of visual complexity and that visual complexity is not a monolithic and static notion but, rather, a multifaceted and dynamic construct. The predictive power of the proposed visual complexity features was examined

using different machine learning algorithms. The results of regression models indicated that the proposed visual complexity features can explain as well as predict videotext difficulty in AL and GA videotext, and in both types of videotext combined. Finally, the chapter introduced AUVANA, an automated tool for extracting, computing, and visualizing visual complexity features in videotexts. It is hoped that the tool will stimulate future research on understanding visual complexity in videotext.

# Chapter 8

# Multimodal Videotext Difficulty Prediction

Previously, in Chapter 6 and Chapter 7, videotext difficulty was modeled as a function of verbal and visual complexity, respectively. The objective was to investigate which modality better explains and predicts difficulty in the SLVC video corpus. The findings, overall, demonstrated that the two unimodal models performed comparably, with verbal models slightly outperforming visual complexity-based models. In this chapter, I investigate whether combining both verbal and visual complexity features into a single predictive model leads to better prediction of videotext difficulty. Particularly, I explored different configurations of ensemble machine learning models and investigated their accuracy in predicting videotext difficulty.

The chapter is organized into three main sections. The first section discusses key approaches and challenges of integrating multimodal data in predictive modeling in the area of Multimodal Machine Learning (MML). The second section introduces the ensemble learning paradigm in machine learning and describes its popular techniques. Finally, the third section presents the main study of this chapter and reports its findings.

## 8.1 Introduction

Multimodal machine learning (MML) is a recent and dynamic area of research in the field of machine learning and pattern recognition. In essence, MML has emerged with an

aim "to build models that can process and relate information from multiple modalities" (Baltrusaitis, Ahuja, & Morency, 2019, p. 423). The word "modality" is often associated with sensory modalities, e.g., vision, touch, smell, and taste, which represent our primary channels of communication and sensation. In the field of MML, a research problem is said to be multimodal when it concerns processing, analyzing, and reasoning about multiple data of such modalities (Baltrusaitis et al., 2019; Guo, Wang, & Wang, 2019).

Based on the heterogeneity of modalities and the relationship between them, multimodal features can fundamentally be classified into three groups:

(a) Heterogeneous multimodal features with no direct relationship between the modalities. For example, eye movement data is a single source and audio data is another.

(b) Heterogeneous multimodal features which describe a single entity. The multimodal features are close in space and time. For example, vocal tone and facial expression.

(c) Multimodal feature describes a single entity using two or more sources of information. While semantically related, the multimodal features can be spatially or temporally distanced.

Many of the problems in machine learning are multimodal in nature (e.g., speech recognition, emotion classification, and multimedia event detection) and it is, therefore, reasonable to assume that integrating multiple dates is more beneficial than relying on a single source of data. Indeed, MML researchers have gained a significant benefit from combining multiple modalities in their work (e.g., Gurban, Thiran, Drugman, & Dutoit, 2008). One of the earliest and most celebrated examples of multimodal machine learning is audio-visual speech recognition systems, where the integration of speech and visual signals has been found to lead to improve speech recognition performance, especially in noisy contexts (Gurban et al., 2008). As digital information is increasingly becoming more multimodal, current unimodal machine learning models are gradually becoming obsolete. For example, a hate classifier trained on user comments only would not be able to detect an image that displays hate or harmful messages.

Despite their great potentials, developing machine learning models that are truly capable of learning from different modalities and reasoning about their relationship is not a trivial task. While humans have no difficulty processing multimodal information and making sense of them, for machines the heterogeneity of information poses many difficulties. In their recent review of the challenges facing the MML research, Baltrusaitis et al. (2019) summarized five key technical challenges:

1. **Representation:** the challenge of representation refers to how to represent the different modalities in a way that exploits the complementarity and redundancy of multiple modalities.

2. **Translation:** the challenge of translation or mapping addresses how to map data from one modality to another, considering that the relationships between modalities are often subjective, and there is no perfect description of such relationships.

3. **Alignment:** the challenge of alignment is concerned with identifying the direct relationship between two or more modalities and resolving issues due to long-range dependencies and ambiguity (e.g., aligning the steps in a recipe to a video showing the dish being made).

4. **Fusion:** the challenge of data fusion refers to how to join information from multiple modalities in a way that is helpful for predictive models.

5. **Co-learning:** the challenge of co-learning refers to how knowledge learned from one modality can be transferred to another learning task, for example, to help develop a model trained on another modality when annotated data are sparse.

While Baltrusaitis et al. presented these five challenges particularly for multimodal deep learning approaches, I believe the challenges are also relevant to traditional multimodal machine learning approaches. I will discuss two of these challenges, the challenge of multimodal representation and multimodal data fusion, as they are more relevant to the current study. It should be noted here that the two challenges are interrelated in the sense that the former is conceptual while the latter is technical.

For developing a multimodal machine learning model, data should be represented in a meaningful format that can help algorithms learn useful information. Good multimodal representations (Bengio, Courville, & Vincent, 2013) or features are therefore the cornerstone for the development of performant multimodal machine learning models (Baltrusaitis et al., 2019). But how to develop multimodal representations?

A straightforward approach is to extract hand-crafted features and fuse them into a single joint representation space. A second approach is to use deep learning for either unimodal or multimodal feature extraction. The two approaches have their advantages and disadvantages. For example, hand-crafted features are easy to extract but may not generalize well to other learning domains. On the other hand, deep learning features are generalizable but extract meaningful multimodal features is difficult and requires large data and expensive computing. In this study, we opted for the first approach for its simplicity.

As far as the challenge of integrating features from different modalities is concerned, three integration methodologies have been proposed in the literature to address this problem; namely, early fusion, late fusion, and intermediate or hybrid fusion (Alpaydin, 2018; Atrey, Hossain, El Saddik, & Kankanhalli, 2010; Poria, Cambria, Bajpai, & Hussain, 2017). In the early fusion approach, features from all modalities are concatenated into a single feature space, and then a machine learning model is trained on the fused features. In the late fusion approach, a separate machine learning model is trained on features from a single modality and independently makes predictions. Then, the predictions from all models are combined—using some mechanisms such as majority vote or averaging—to make a final decision. Lastly, the intermediate or hybrid fusion approach lies between the two extremes and combines the two approaches into a single process. Specifically, in this hybrid approach, each modality is initially processed to construct or generate an abstract representation and then the generated representations from all modalities are fed into a single machine learning algorithm (Alpaydin, 2018). Each of the three fusing approaches has its strengths and limitations and in practice, each approach works best with a particular type of task or problem (Atrey et al., 2010). It is more appropriate, for example, to use an early integration when the feature space is relatively small, and the features are independent. In this study, I used both early and late fusion techniques for joining complexity features from different modalities.

## 8.2 Ensemble Learning

Ensemble learning is a machine learning paradigm where multiple machine learning models (commonly called "weak learners" or "base models") are combined to get better predictive accuracy (Witten et al., 2016). The main hypothesis behind this approach is that when several "weak" learners are correctly combined to solve the same problem they often lead to better accuracy than a single model (Z. Zhou, 2012). This is because an ensemble model is more capable of reducing prediction error while also being more complex.

There are several ways for developing ensemble models. Common ensemble methods include averaging, bagging, boosting, and stacking, each of which has several variants in the literature. in essence, ensemble methods differ in two main ways. First, whether the aggregated base models are "heterogeneous" or "homogeneous." An ensemble model is said to be homogeneous when it consists of members having a single-type base learning algorithm (e.g., bagging). On the other hand, a heterogeneous ensemble model consists of members having different base learning algorithms. The second distinction between

| Model | Type of features | Fusion | Ensemble |
|---|---|---|---|
| OLS regression | Multimodal | Ealry | No |
| Weighted Average Regressor | Uni-modal | late | yes |
| Stacked Regressor | Uni-modal | late | yes |
| Random Forest | Multimodal | Early | yes |
| Gradient Boosting | Multimodal | Early | yes |
| Adaboost | Multimodal | Early | yes |
| Stacking | Multimodal | Early | yes |

ensemble techniques is whether the learning algorithms are fitted to the dataset in a "parallel" or "sequential" fashion (see Z. Zhou, 2012, for a comprehensive review of ensemble learning techniques).

In what follows, I briefly describe the four ensemble learning techniques used in this study and highlight their key differences in the context of regression problem:

1. **Averaging:** Averaging is a simple ensemble technique that can be used for both regression or classification problems. In this approach, multiple base models (often equally performing models) are trained independently (in a parallel fashion) on the entire training set. The predictions made by these base models on the testing set are then averaged to make the final prediction. While averaging helps in balancing out the weaknesses of the individual models, a low-performing model can drag down the overall performance of the ensemble models. One solution to this problem is to remove the low-performing model. Another solution is to add some weights to reward base models that perform well and penalize those that perform poorly. This approach is called the weighted average ensemble learning technique.

2. **Bagging:** Bagging stands for bootstrap aggregation. In this ensemble learning approach, multiple weak learners (often homogeneous learners) are trained on different subsets of the dataset using bootstrap sampling—a subset of the dataset is chosen randomly with replacement—and then the predictions of base models are aggregated to generate the final prediction. An example of a bagging meta-algorithm is a random forest that combines individual decision trees.

3. **Boosting:** Boosting[1] is a sequential method that aims to "convert weak learners to strong learners" (Z. Zhou, 2012, p. 23). Like bagging, boosting methods combine several uncorrelated weak learners for generating a strong model. However, unlike bagging, boosting methods consist of fitting weak learners sequentially so that each

[1]Boosting technique is very similar to additive models in statistics (Witten et al., 2016; Z. Zhou, 2012)

newly fitted model focuses its effort on handling observations that were difficult for the previously fitted model. Two popular machine learning algorithms using this technique are gradient boosting and adaptive boosting (AdaBoost).

4. **Stacking:** in the stacking ensemble technique, multiple base models are trained on the dataset and the predictions of base-level models on the test set are used as an input (features) in a meta-regressor. The base-level can consist of different learning algorithms and hence stacking ensembles are often heterogeneous.

The use of ensemble learning generally helps in mitigating the bias and variance problem and it also often leads to an increased prediction accuracy (Witten et al., 2016; Z. Zhou, 2012). However, there are several shortcomings of ensemble learning. First, ensemble models lack interpretability. This is because when combining dozens or hundreds of individuals, it becomes more difficult to determine what factors are contributing to the improved decisions (Witten et al., 2016). Another shortcoming of ensemble learning is that it is computationally expensive. This is especially true when the ensemble model consists of hundreds or thousands of complex base learners (e.g., neural nets) or it relies on sequential learning.

## 8.3 Statistical Procedure and Analysis

In this section, I first outline the study's research questions. Then, I describe the methodology implemented to address them. Secondly, I present the study findings.

### 8.3.1 Research Questions

The study reported in this chapter addresses the following three research questions:

1. Do multimodal regression models lead to better prediction of videotext difficulty than unimodal models developed in the previous two chapters?
2. Does the use of a stacked model with base models trained independently on each modality lead to better prediction of videotext difficulty?
3. Does the use of ensemble learning methods (averaging, bagging, boosting, and stacking) lead to better prediction performance than a single model? And if so, what ensemble approach is the most predictive of videotext difficulty?

### 8.3.2 Corpus

For the current study, the entire SLVC corpus was used for developing genre-agnostic predictive models, and the two video subcorpora, AL and GA, were used for developing predictive models for each video genre.

### 8.3.3 Method

In this study, we explored three different setups or configurations of machine learning algorithms and videotext complexity features. The three configurations are described below:

#### 8.3.3.1 Configuration One

A simple OLS regression model was developed for each of the three video corpora (AL, GA, and ALL) using multimodal complexity features. The configuration is visualized in Figure 8.3.

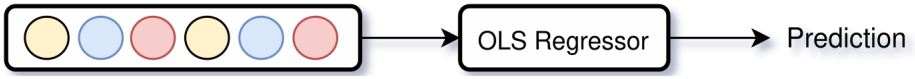



FIGURE 8.1: A Visual Representation of Three Configurations

The main objective of this setup is to develop a parsimonious model. As by the principle of parsimony, when considering multiple competing models, one should seek the simplest model so that fewer parameters need to be estimated. This configuration addresses the first research question, which investigates if using multimodal complexity features leads to better prediction of videotext difficulty than using unimodal features (verbal or visual complexity features alone). Furthermore, the developed model will be used as a baseline model for comparing the performance of the more complex models in configurations two and three.

The predictive powers of multimodal models are then compared to the models developed using either verbal or visual complexity features alone.

#### 8.3.3.2 Configuration Two

When the feature set is large (as in our case), there is always a risk of having spurious relationships among the features. Therefore, in configuration two, we address this issue by developing three ensemble models (using stacking, averaging, and weighted average techniques) with three base regression models that were independently developed on complexity features extracted from each of the three modalities. The three ensemble models vary on how the final prediction is made. In the stacking model, the predictions from each base model are used as input for a successive meta regressor which makes the final prediction.

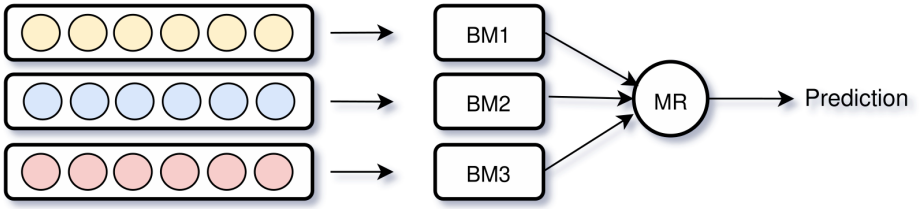



FIGURE 8.2: A Visual Representation of Unimodal Ensemble Model Configuration

We hypothesized that through using this configuration, we can strike a balance between ensemble model interpretability and its prediction ability.

#### 8.3.3.3 Configuration Three

Ensemble multimodal models of several base models trained on multimodal complexity features using different meta-learning approaches: boosting, bagging, and stacking.

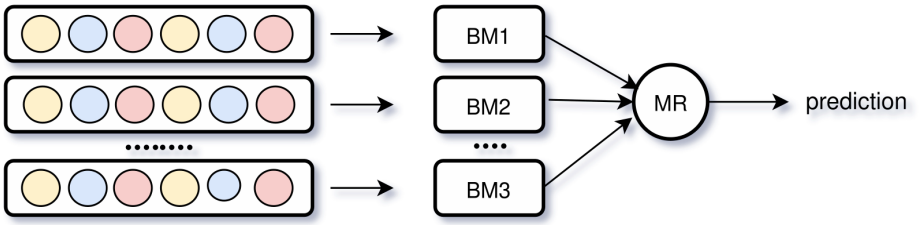



FIGURE 8.3: A Visual Representation of Multimodal Ensemble Model Configurations

### 8.3.4 Feature Selection

In this study, an aggregation of the previously extracted verbal complexity features (Chapter 6) and visual complexity features (Chapter 7) made the multimodal complexity features set. In total, there were 290 multimodal features. The features were early fused or concatenated to form a single compact feature space. As discussed earlier, one drawback of this approach is that when combining features from all modalities the feature space becomes high-dimensional. This may cause three problems: 1) processing a large number of features means a higher compactional cost, 2) three is a risk of overfitting, and 3) because multimodal features extracted from different modalities, their units and scales are different (Alpaydin, 2018).

To reduce the number of features and prevent overfitting, I followed the same feature selection routine used previously in Chapter 6 and Chapter 7. Using this routine on the entire video dataset, a subset of 31 were selected by a LASSO regressor. The number of features was further reduced to 10 features using backward elimination to remove insignificant predictors. The same procedures were also implemented on AL and GA datasets, resulting in 9 and 13 multimodal complexity features, for predicting the difficulty in each video genre, respectively.

## 8.4 Results

### 8.4.1 Configuration One

#### 8.4.1.1 Multimodal Regression Models: ALL Video Corpus

A genre-agnostic regression model was developed using 10 multimodal complexity features selected using our feature selection routine. The multimodal features included in the model were: *Std pitch, AWL Sub-list 3 Normed, All AFL Normed, Root TTR (AW), MRC Imageability AW, COCA spoken tri 2 AC, CN C, Keyframe KB (mean), SSI Video W5Sec (SD)*, and *Video H264 Subsample Size (MB)*. Table 8.1 shows the model coefficients and additional statistics.

The performance of the OLS regression model was cross-validated and the finding demonstrated that the trained model yielded an RMSE of 2.75. The 10 visual complexity features

Table 8.1: Summary for Multiple Regression Model for Predicting Videotext Difficulty Using Multimodal Features

| Features | $\beta$ | SE | $t$ | Sig. |
|---|---|---|---|---|
| Pitch (SD) | -0.51 | 0.15 | -3.44 | <.001 |
| AWL Sub-list 3 (Normalized) | 0.47 | 0.14 | 3.3 | <.001 |
| All AFL Normed | -0.39 | 0.16 | -2.38 | .024 |
| Root TTR (AW) | 1.00 | 0.16 | 6.4 | <.001 |
| MRC Imageability AW | -0.59 | 0.17 | -3.54 | <.001 |
| COCA spoken tri 2 AC | 0.33 | 0.14 | 2.35 | .027 |
| Complex Nominal per Clause | 0.57 | 0.16 | 3.64 | <.001 |
| Keyframe KB (mean) | -0.4 | 0.15 | -2.68 | .011 |
| SSI Video W5Sec (SD) | 0.58 | 0.15 | 3.96 | <.001 |
| Video H264 Subsample Size (MB) | 0.67 | 0.16 | 4.19 | <.001 |

*Note.* Constant term= 16.83; $\beta$ is the standardized beta coefficient; SE is standard error

collectively explained 38% of the variance in the testing set. The Pearson correlation between the true and predicted videotext complexity scores were $r = .54, p < .001$.

#### 8.4.1.2 Multimodal Regression Models: AL Video Corpus

The regression model developed for predicting the difficulty of AL videotexts achieved an RMSE of 2.5. The model was developed using nine multimodal complexity features: *Maas TTR (CW), Kuperman AoA, CW, COCA spoken bigram T, COCA academic trigram 2 T, Determiners, Adjacent overlap noun sent div seg, Number of Moving Objects per minute, Motion Duration Difference (Variance)*, and *Number of shots*. Table 8.2 shows the AL model coefficients along their significant values.

The results have also shown that the developed AL regression model explained about 51% of the variance. The correlation of the model prediction and the actual rating of videotext difficulty was $r = .72, p < .001$, indicating the model predicted a large portion of the samples correctly.

#### 8.4.1.3 Multimodal Regression Models: GA Video Corpus

Using the feature selection routine described earlier, a set of 13 multimodal complexity features was selected to be good predictors of difficulty in GA videotext. The features were: *k2 percent, AWL Sub-list 3 Normed, Lexical density token, MTLD ma wrap (CW),*

Table 8.2: Summary for Regression Model for Predicting AL Videotext Difficulty Using Multimodal Features

| Features | $\beta$ | SE | $t$ | Sig. |
|---|---|---|---|---|
| Maas TTR (CW) | -0.75 | 0.24 | -3.07 | <.001 |
| Kuperman AoA (CW) | 1.44 | 0.22 | 6.46 | <.001 |
| COCA spoken bigram T | 0.53 | 0.18 | 2.91 | <.001 |
| COCA academic trigram 2 T | -0.59 | 0.19 | -3.16 | <.001 |
| Determiners | -0.53 | 0.18 | -2.9 | <.001 |
| Adjacent overlap noun sent div seg | -0.48 | 0.17 | -2.78 | .012 |
| Number of Moving Objects per minute | 0.42 | 0.19 | 2.22 | .039 |
| Motion Duration Difference (Variance) | 0.76 | 0.19 | 3.95 | <.001 |
| Number of shots | 0.53 | 0.22 | 2.43 | .024 |

*Note.* Constant term= 17.23; $\beta$ is the standardized beta coefficient; SE is standard error

*Kuperman AoA CW, COCA spoken tri MI, Positive causal, Positive logical, CN C, Colorfulness Adjacent KF (mean), Colorfulness Adjacent Video W5Sec (SD), Video Compression H265 Size (MB)*, and *Number of Faces per Minute*. Table 8.3 shows the model coefficients along with other useful statistics.

Table 8.3: Summary for Regression Model for Predicting GA Videotext Difficulty Using Multimodal Features

| Features | $\beta$ | SE | $t$ | Sig. |
|---|---|---|---|---|
| K2 COCA-BNC | -0.72 | 0.22 | -3.2 | <.001 |
| AWL Sub-list 3 Normed | 0.62 | 0.21 | 2.93 | <.001 |
| Lexical density token | -0.6 | 0.20 | -2.96 | <.001 |
| MTLD ma wrap (CW) | 0.67 | 0.22 | 3.11 | <.001 |
| Kuperman AoA CW | 0.72 | 0.25 | 2.87 | .016 |
| COCA spoken tri MI | -0.62 | 0.2 | -3.11 | <.001 |
| Positive causal | -0.55 | 0.21 | -2.65 | .012 |
| Positive logical | 0.79 | 0.21 | 3.79 | <.001 |
| Complex Nominals per Clause | 0.84 | 0.25 | 3.36 | <.001 |
| Colorfulness Adjacent KF (mean) | -0.8 | 0.23 | -3.54 | <.001 |
| Colorfulness Adjacent Video W5Sec (SD) | 0.66 | 0.22 | 2.95 | <.001 |
| Video Compression H265 Size (MB) | 0.6 | 0.21 | 2.86 | .019 |
| Number of Faces per Minute | -0.67 | 0.20 | -3.34 | <.001 |

*Note.* Constant term= 16.92; $\beta$ is the standardized beta coefficient; SE is standard error

To assess its predictive power and generalizability, the model was cross-validated, the results showed the model has an RMSE of 2.50 its predictions correlated significantly,

$r = .50, p < .001$, with the true difficulty values of videos in this genre. The model explained 28 % of the variance in the dataset.

#### 8.4.1.4 Comparison of Unimodal vs. Multimodal Models

A comparison of the prediction performances of the verbal-based models, visual-based models, and multimodal models is shown in Table 8.4. The results showed that multimodal models outperformed the verbal-only and visual-only models in predicting the difficulty of videotexts in the three video datasets (AL, GA, and ALL). This finding suggests that the integration of verbal and visual complexity features in a single regression model is more predictive of videotext difficulty than using either verbal or visual complexity alone.

Table 8.4: The Prediction Performance (10-Fold Cross-Validation) of Unimodal Models vs. Multimodal Models

| Models | AL | | GA | | ALL | |
|---|---|---|---|---|---|---|
| | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ |
| OLS regression model (Verbal) | **2.47** | **.51** | 2.83 | .25 | 2.75 | 33 |
| OLS regression model (Visual) | 2.71 | .45 | 2.94 | 22 | 2.90 | .29 |
| OLS regression (Multimodal) | 2.5 | .51 | 2.84 | .28 | 2.75 | .38 |
| Random forest model (Multimodal) | 2.73 | .46 | 2.91 | .17 | 2.74 | .38 |
| Gradient boosting model (Multimodal) | 2.84 | .42 | 2.93 | .21 | 2.87 | .31 |
| Adaboost model (Multimodal) | 2.69 | .44 | 3.13 | .17 | 2.72 | .36 |
| Stacking model (Multimodal) | **2.4** | **.55** | **2.74** | **.29** | **2.64** | **.42** |

### 8.4.2 Configuration Two: Ensemble Unimodal Models

The objective of this configuration is to investigate whether stacking three models trained independently on acoustic, linguistic, and visual complexity features can lead to an increase in the predictive power of the model. Three OLS regression models were developed using a subset of complexity features from each of three modalities (acoustic, linguistic, and visual). The predictions from each of the three independent models were used as input to meta-regressor or estimator (here a ridge estimator) to make the final prediction.

Using this configuration, I developed a stacked ensemble model for each video genre, AL and GA, as well as for both video genres combined. The performances of the multimodal models were validated using a 10-fold cross-validation approach. Table 8.5 shows the

RMSE of three stacked models, the coefficients of determination ($R^2$), and the correlation between models' prediction and actual values of the dependent variable.

Table 8.5: A Comparison of the 10-fold Cross-Validation Performances of Ensemble Unimodal Models on Predicting the Difficulty in AL, GA, and ALL Video Corpus

| **Models** | **AL** | | **GA** | | **ALL** | |
|---|---|---|---|---|---|---|
| | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ |
| Stacked unimodal model | 2.54 | .45 | 3.19 | .19 | 2.59 | .43 |
| Weighted average unimodal model | 2.41 | .55 | 2.79 | .27 | **2.67** | .41 |

The results demonstrated that the three stacked unimodal regression models did not outperform the OLS regression models in configuration one. One possible explanation for this finding is that the three unimodal models were not equally important in terms of predicting video difficulty.

The weighted average ensemble technique allows multiple base models to contribute to the prediction proportional to their estimated performance or how well they are performing in the prediction task. It is an extension of the simple average ensemble technique where the predictions of base models are simply averaged without rewarding well-performing base models. In practice, the weighted average model performs as well as the best model in the ensemble. To develop the weighted average model, three OLS regression models were initially developed using a set of strong predictors extracted from each modality than a vector of weights $\vec{w}$ was added manually to three models based on their performance. Based on an experimentation with different weights, the best weight vector was $\vec{w} = [0.1, 0.5, 0.4]$ corresponding to acoustic, linguistic, and visual-based model, respectively. That is, the linguistic base model contributed the most to the prediction ability of the ensemble model, followed by the visual base model and the acoustic base model.

### 8.4.3 Configuration Three: Ensemble Multimodal Models

In configuration three, I developed several ensemble models using multimodal features and a combination of different machine learning algorithms and ensemble techniques. Specifically, I used Random Forest Regressor (RFR), gradient boosting regressor, and AdaBoost Regressor. The hyperparameters for all algorithms were tuned using a randomized grid search technique which randomly samples from a predefined grid of hyperparameters and performs cross-validation with each combination of values. Hyperparameter

tuning is a very important step for not only making learning more efficient but also for ensuring that the model does not overfit.

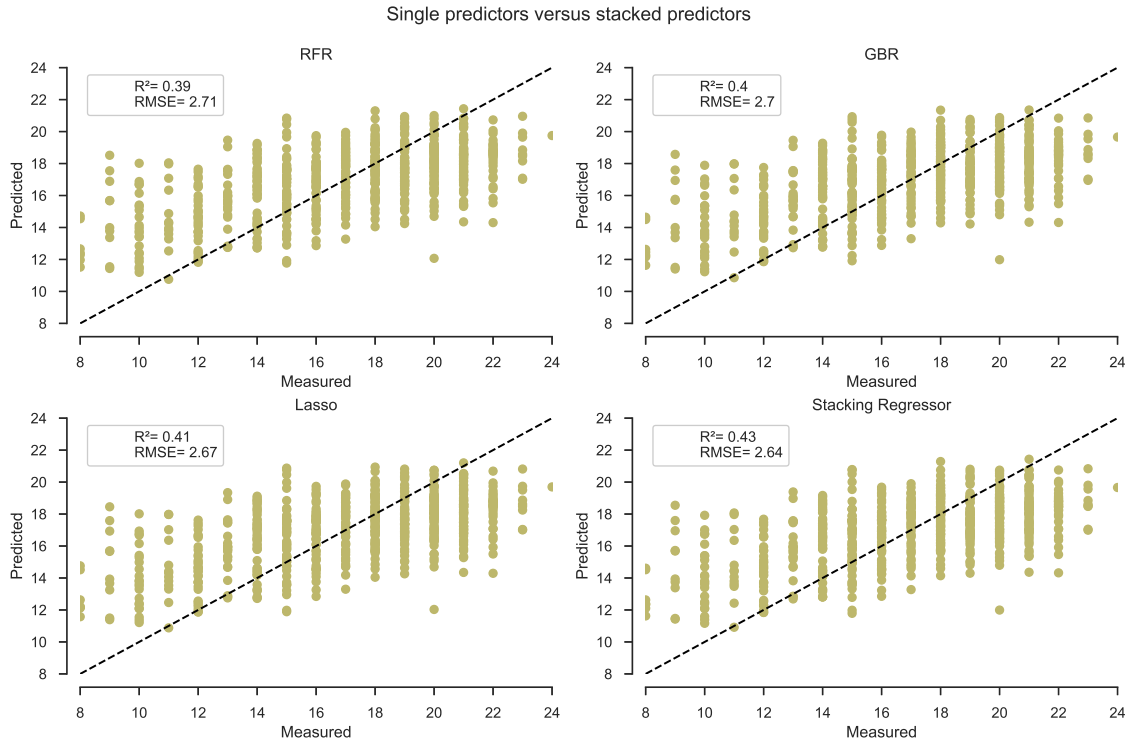



FIGURE 8.4: The Performances of Single Base Models vs. a Stacked Model

I also developed a stacked model using the tuned RFR, GBR, and LASSO algorithms as base models and a simple OLS algorithm as a final meta-learner. The predictive performances of all ensemble models were validated using a 10-fold cross-validation approach. For comparing the performance of the ensemble models, RMSE was used. The prediction performances of the models are shown in Table 8.6.

TABLE 8.6: A Comparison of The 10-Fold Cross-Validation Performances of Ensemble Models Predicting The Difficulty In AL, GA, And ALL Video Corpus

| **Models** | **AL** | | **GA** | | **ALL** | |
|---|---|---|---|---|---|---|
| | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ |
| Random forest model | 2.73 | .46 | 2.91 | .17 | 2.74 | .38 |
| Gradient boosting model | 2.84 | .42 | 2.93 | .21 | 2.87 | .31 |
| Adaboost model | 2.69 | .44 | 3.13 | .17 | 2.72 | .36 |
| Stacking model | **2.4** | .55 | **2.74** | .29 | **2.64** | .42 |

## 8.5 Discussion

The main objective of the chapter was to investigate whether combining verbal and visual complexity features into predictive multimodal models leads to an improved prediction accuracy of videotext difficulty. To this end, I explored several possible configurations of integrating multimodal complexity features into a single ensemble machine learning model. These configurations allowed us to explore several predictive models that vary in their interpretability and complexity.

The first research question asked whether combining verbal and visual features in a single regression model leads to better prediction of videotext than developing models using verbal or visual features alone. The results showed that multimodal models are somewhat better than uni-modal models in predicting the difficulty of videotext in the SLVC corpus. Concerning the contributions of each modality on videotext complexity, lexical, syntactic, and visual complexity features were the top predictive features in the three models, whereas the acoustic complexity features were the least predictive. In particular, a single acoustic feature (Pitch SD) was induced in the regression model developed for predicting videotext difficulty in both video genres, and no acoustic features were included in AL and GA models. This result seems to indicate that for intermediate language learners acoustic and phonological complexity features are less predictive of videotext difficulty in comparison to lexical, syntactic, and visual complexity features. One possible interpretation of this finding is that the multimodal nature of videos may have reduced the difficulty associated with acoustics.

The second question was concerned with investigating the predictive performance of the unimodal ensemble model. The hypothesis behind this configuration was to leverage the contributions of strong unimodal models while also maintaining high model interpretability. The results generally showed that both stacked unimodal models and weighted average unimodal models outperformed simple OLS multimodal regression models (in configuration one), with the latter type of models, are showing. This finding indicates that three modalities (sound, language, and visual) are contributing differently to predicting L2 videotext difficulty.

The last question addressed in this study investigated whether the use of more advanced ensemble learning techniques and multimodal complexity features lead to better prediction of videotext difficulty compared to models developed in the first and second configurations. The results showed that a stacking model with three tuned base learners and

an OLS as a meta-learner had the highest prediction accuracy, though it only slightly outperformed the best-performing unimodal model.

### 8.5.1 Contributions and Limitations

The study reported in this chapter was innovative in several regards. To the best of my knowledge, the current study was the first investigation into the impact of multimodal complexity features in predicting videotext difficulty for L2 language learners. The study further explored different ensemble learning schemes and whether they help in boosting the predictive performance of multimodal complexity models. As the results of the study have shown all ensemble models had higher predictive accuracy than simple OLS regression models. Therefore, when the goal of videotext difficulty modeling is to accurately predict videotext difficulty, the finding of the current study suggests the use of ensemble models for predicting videotext difficulty.

However, several limitations should be acknowledged. In developing multimodal videotext difficulty models, we simply concatenated features from three modalities without considering the interaction between the modalities. There is a bi-directional relationship between linguistic and visual processing; language disambiguates visual processing and vice versa (Coco, Keller, & Malcolm, 2016). Future studies should consider accounting for the interaction between the different modalities in videotext and how their interaction may or may not contribute to language learners' perception of videotext difficulty.

Another possible avenue for future research is investigating whether or not the use of features extracted from deep learning networks is better than hand-crafted features in predicting videotext difficulty. Recently, deep multimodal learning representation has shown great results in video classification (e.g., Y. Liu, Feng, & Zhou, 2016). However, it is still unclear if using deep learning-based features would be beneficial for videotext difficulty assessment.

## 8.6 Conclusion

In this chapter, I investigated whether the use of multimodal complexity features improves the prediction of videotext difficulty. I also investigated the impact of developing ensemble models using different techniques (weighted average, boosting, bagging, and

stacking) in predicting videotext difficulty. The results overall demonstrated that firstly the use of multimodal complexity features leads to better prediction of videotext difficulty than using unimodal complexity features. Secondly, the use of multimodal ensemble predictive models outperformed the simple multimodal regression model and unimodal ensemble regression models in accurately predicting the difficulty of AL, GA, and ALL videotexts.

# Chapter 9

# Real Time Analysis and Forecasting of Video Segment Difficulty

In the previous three Chapters (6, 7, and 8), I investigated the contribution of verbal, visual, and multimodal complexity on predicting the difficulty of the entire videotext. While useful, a post-hoc estimation of videotext difficulty forfeits the chance to capture more nuanced insights of difficulty in videotext. In this chapter, I propose an analytical approach for analyzing and forecasting the difficulty of individual segments of videotext using univariate time series forecasting. To investigate the efficacy of this approach, several sequential neural network models were trained and validated on a corpus of 34,363 short video segments (5 seconds long each). Finally, I conducted a qualitative analysis to examine the trained model's accuracy in forecasting the difficulty of unseen videos.

Structurally, the chapter is organized into three sections. The first section motivates the need for an online, real-time assessment of videotext difficulty in the context of language learning and assessment. presents the study's research questions. The second section describes the basic properties of time series data and discusses the use of artificial neural networks—Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRUs)—in univariate time series analysis and forecasting. The third section presents our unsupervised approach for developing a multistage system for forecasting the difficulty of short video segments.

## 9.1 Introduction

Videotext understanding is an incremental process (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Viewers continuously assign meaning to the information coming from the different video modalities and update their mental models using current and prior knowledge (Venhuizen, Crocker, & Brouwer, 2019). Amidst this process, it is very plausible that breakdowns in comprehension occur, and the viewer needs to remedy those breakdowns as quickly as possible. However, to date, little, if any, a systematic effort has been made to analyze and forecast the difficulty of short segments of videos in real-time (but see Srivastava, Velloso, Lodge, Erfani, & Bailey, 2019).

It is assumed that when watching a video lecture, say in biology, a language learner may find certain parts of a video lecture more difficult than others. Identifying those difficult parts of the videotext therefore can be useful for several applications, for example, in determining where to split a long instructional video or where to embed activities in an interactive video learning environment. Another possible application is in video-based language assessment, where different scoring weights can be assigned to questions related to more difficult video segments. Test writers can also use automated video difficulty assessment systems to identify where they might target items in videos.

Besides practical applications, real-time video difficulty assessment can help researchers to design better and more informed studies and pave the way to the development of theories of L2 videotext comprehension and their computational instantiations thereof.

In this chapter, I propose an unsupervised approach for forecasting difficulty in real-time. Specifically, the problem of forecasting video segment difficulty is treated as the problem of detecting an anomaly in univariate time series data. In the sections that follow, I first define what time series data are and explicate their fundamental proprieties. Next, I review and discuss the use of neural network approaches for forecasting anomalies in time series data. Finally, I present the study's questions, describe the approach developed in this study, and show the results of the study.

### 9.1.1 Time Series Data and Their Basic Proprieties

In essence, time-series data "is a collection of observations taken sequentially in time" (Chatfield, 2000, p. 1). The analysis and modeling of time series data have a long history in fields such as economics, finance, and business and it has been proved useful for numerous

tasks in many other branches of sciences. Examples of time series data are countless, including the stock market, weather readings, speech, text, video data, and brain activities recorded from an EEG machine. The social sciences also furnish many examples of time series and latent growth modeling is a case in point.

What makes time series data different from other types of data is that there is an explicit chronological order dependence between observations. Depending on the objective of time series modeling, a time series can be analyzed for describing patterns in the data or forecasting future events based on understanding past events. For each goal, some techniques are particularly more useful than others, hence determining the objective of the analysis beforehand can alleviate a large technical investment in time and expertise that are not directly aligned with the desired outcome. In forecasting, it is important to formulate the problem carefully and how and for what purpose the forecasting system will be used.

Many of the forecasting scenarios—and indeed many of the books on the topic—are based on univariate time series analysis. Where in univariate time series analysis a single dependent variable is being recorded or observed at equally spaced time points, in multivariate time analysis there are multiple time-dependent variables. There are several objectives of doing a multivariate time analysis, for example, to study the dynamic relationships between the variables and to improve the accuracy of the forecasting (Tsay, 2013). In time series forecasting, it is rarely assumed that variables keep unchanging. What is normally assumed is that how the time series is changing continues (Hyndman & Athanasopoulos, 2018). A good forecasting model, therefore, captures historical changes in the past and accurately forecasts them in the future.

### 9.1.2 Deep Learning Time Series Forecasting

The use of neural network-based approaches for time series analysis and forecasting has recently shown a remarkable leap in prediction performance compared to traditional, statistical approaches. Particularly, recreant neural networks (RNNs) and their variants, especially gated RNNs, have been useful for modeling time series data and generating more accurate forecasting.

One of the advantages of using Artificial Neural Networks (ANNs) for modeling time series data is that they do not presume many of the assumptions in statistical time series analysis approaches such as moving average (MA), autoregressive–moving-average

model(ARMA), or ARIMA. In particular, ANN approaches do not require time-series data to follow a particular known statistical distribution (e.g., the Gaussian distribution) (Adhikari & Agrawal, 2013) or the existence of stationarity, seasonality, and autocorrelation in the time series. This is because of the flexibility of ANNs to learn nonlinear patterns in the data (Hornik, Stinchcombe, & White, 1989). Having said that, ANNs lack the interpretability that classical approaches afford.

In this study, I trained several neural networks for the task of forecasting the difficulty in unseen video segments. In what follows, I briefly describe the architectures of both neural networks and highlight their capabilities for time series forecasting.

#### 9.1.2.1 Recount Neural Networks (RNNs)

RNNs are a class of artifactual neural networks that are particularly suitable for modeling sequential data. RNNs differ from traditional feedforward neural networks in that they permit input. This dynamic behavior makes RNNs useful for modeling sequential data and thus, they are commonly used in fields such as natural language processing, audio analysis and speech recognition, and text translation. Figure 9.1 compares the generic architecture of the feedforward ANN and RNNs.

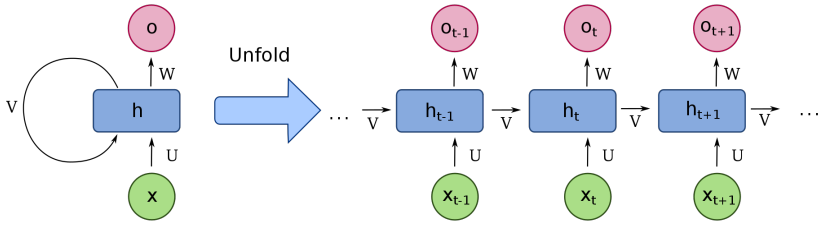



FIGURE 9.1: A Standard Recurrent Neural Networks (RNNs)

#### 9.1.2.2 Long Short-Term Memory (LSTM)

The LSTM neural network architecture is an extension of the Recurrent Neural Networks (RNNs) (Hochreiter & Schmidhuber, 1997). The LSTM network was purposefully designed to tackle the gradient vanishing or exploding problem in RNNs (Pascanu, Mikolov, & Bengio, 2013). In other words, RNNs, in practice, fail to establish long-term dependencies (Pascanu et al., 2013). To solve this problem, the LSTM unit contains three gates that

can allow or block information from passing by (see Figure 9.2). The three gates are commonly called the forget gate, input gate, and output gate. Each of the three gates has a specialized function to perform in the LSTM unit. All three gates consist of a sigmoid neural net layer along with a pointwise multiplication operation. The sigmoid function output ranges from 0 to 1, where 0 indicates no information should be allowed to pass to the next cell and 1 indicates all information should be passed to the next cell in the network.

$$
\begin{aligned}
i_t &= \sigma(x_t U^i + h_{t1} W^i) \\
f_t &= \sigma(x_t U^f + h_{t1} W^f) \\
o_t &= \sigma(x_t U^o + h_{t1} W^o) \\
\tilde{C}_t &= \tanh(x_t U^g + h_{t1} W^g) \\
C_t &= \sigma(f_t * C_{t1} + i_t * \tilde{C}_t) \\
h_t &= \tanh(C_t) * o_t
\end{aligned}
\tag{9.1}
$$

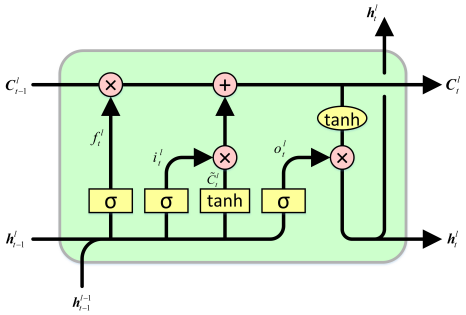



FIGURE 9.2: A Long Short-Term Memory (LSTM) Unit

#### 9.1.2.3 Gated Recurrent Unit (GAU)

Similar to LSTM, the gated recreant unit network (Cho et al., 2014) also addresses the problem of the short-term dependencies in RNNs. However, unlike the LSTM unit, the GAU uses two gates instead of three. These two gates also control the flow of information in the network, but they do so in a more efficient way. Figure 9.3 shows a single GAU.

It combines the forget and input gates into a single "update gate." It also merges the cell state and hidden state and makes some other changes. The resulting model is simpler than standard LSTM models and has been growing increasingly popular.

$$
\begin{aligned}
z_t &= \sigma(W_z \cdot [h_{t1}, x_{tt}]) \\
r_t &= \sigma(W_r \cdot [h_{t1}, x_t]) \\
\tilde{h}_t &= \tanh(W \cdot [r_t * h_{t1}, x_t]) \\
h_t &= (1 - z_t) * h_{t1} + z_t * \tilde{h}_t
\end{aligned}
\tag{9.2}
$$

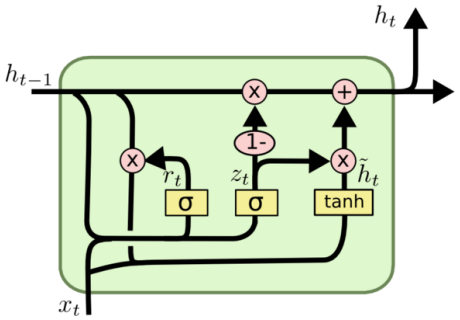



FIGURE 9.3: A Gated Recurrent Unit

### 9.1.3 Time Series Models Evaluation

When estimating the predictive performance of a machine learning model, the k-fold cross-validation (CV) approach is more robust and reliable than other approaches. However, using this approach is not directly applicable for time series data because of the time dependency between the variables. More importantly, when training a model on sequential data, our objective is that our model learned the dependency among the features in the data, therefore breaking the order of the data may not help the model to learn and consequently lead to poor generalizability and prediction accuracy. Therefore, we need a way to accurately evaluate the trained model while also preserve the temporal order of the observations in the dataset. In time series analysis literature, there are several performance evaluation techniques proposed to address the dependency issue in time-series data, which fundamentally are different variants of two general methodologies; k-fold cross-validation (CV) and out of sample (OOS) (Bergmeir, Hyndman, & Koo, 2018; Cerqueira, Torgo, & Mozetic, 2019). A key distinction between the two methodologies is that the latter preserves the temporal order of the observations and never uses past data

for model evaluation (Cerqueira et al., 2019). And in terms of computational speed, the out-of-sample approach is faster than the cross-validation approach. However, when the data is limited, the k-fold evaluation approach is more preferred as it makes efficient use of the data (Bergmeir et al., 2018). In a recent study, Cerqueira et al. (2019) conducted an extensive experiment of the two evaluation strategies on different types of data and time series scenarios and reported that out-of-sample evaluation methods are more suitable in most cases, especially when there is a variation in the time series data because of non-stationarity.

## 9.2 Method

The main objective of the study is to propose an analytical approach for forecasting the difficulty of video segments using artificial neural networks. I use the term "forecasting" here to denote the task of predicting the next value of the time series, $yt + 1$, given the previous observations of $Y$. addressed in this chapter is that giving the multimodal complexity features and difficulty of past video segments, can we forecast the difficulty of the next video segment. Specifically, the current study addresses the following two research questions:

### 9.2.1 Research Questions

1. Given a global rating of videotext difficulty; how can we predict the difficulty of short video segments in the videotext?

2. How accurately can we forecast the estimated difficulty of video segments using artificial neural networks?

### 9.2.2 Corpus

Because of the computational cost of extracting and computing multimodal complexity features at a very short time interval, in this study, I only considered training a genre-agnostic model for forecasting the difficulty of video segments to validate our approach. For model development, I used the entire SLVC corpus (n=640 videotexts).

### 9.2.3 Our Approach

As discussed earlier, assessing the difficulty of a short video segment, as opposed to an entire videotext, can be useful for more in-depth analysis and real-time assessment of videotext difficulty. However, obtaining an accurate estimate of a short video segment (e.g., 5 or 10 seconds long) is a very laborious and expensive task. Therefore, to address this issue I propose an analytical approach for forecasting difficulty that relies on the judgment of the difficulty of the entire video to obtain predicted difficulty values for shorter video segments. The approach is implemented in three steps:

- Develop a performant OLS linear regression model using a representative set of video-level multimodal complexity features.
- Use the coefficients of the regression model (from stage 1) to generate an estimated prediction of difficulty for each video segment.
- Train a deep neural network to forecast video segment difficulty on unseen videotext.

The main goal of Stage 1 and 2 is to make an estimated difficulty score for each video segment which then becomes the ground truth values to be predicted in Stage 3. In the sections that follow, I describe the implementation of these three steps and discuss some issues and considerations taken at each step. Then, I present the validation result of the forecasting model. To avoid confusion, I shall refer to the video-level model developed in step 1 as "model one", and the forecasting model developed in step 3 as "model two".

#### 9.2.3.1 Stage One: A Video-Level Regression Model

Before implementing this step, we need first to decide on what is the most appropriate time interval or window size of a video segment. It should be noted that there is a trade-off between the window size of the video segment and the amount of information retained. That is, with larger window sizes, more information is kept for analysis and modeling, which in turn helps in increasing the prediction accuracy of the model. However, a smaller window size allows for a fine-grained analysis of video data but may result in a poor-performing model. The decision also helps in determining what multimodal complexity features to consider in the video-level regression model. Because we are interested in estimating video difficulty at short time intervals (for real-time forecasting of

video difficulty), the window size of video segments should therefore be small. Moreover, a shorter video segment allows for aggregating data or downsampling the length of the video segment.

To investigate the effect of window size on prediction accuracy, I developed two regression models to predict video difficulty using only multimodal features that can be computed (or extracted) at two different window sizes, 5 and 10 seconds. After disregarding features that cannot be computed within these time windows, such as syntactic, discoursal, and some visual complexity features, the remaining multimodal complexity features (n=136) were used in developing the two models.

For determining the features to be included in the two models, I followed the same feature selection routine used in the previous chapters. After fitting the models, the prediction performances of the two developed models were validated using the 10-fold cross-validation technique. The results demonstrated that the two regression models (five-second and ten-second) yielded comparable results: RMSE: 2.59 vs. 2.57, respectively. Therefore, the five-second model (henceforth referred to as the video-level regression model) was used for generating an estimated difficulty score for each video segment in the SLVC corpus. Table 9.1 shows the coefficients of the video-level regression model.

Table 9.1: Summary for Level-one Model for Estimating Video Segment Difficulty

| Features | $\beta$ | SE | $t$ | Sig. |
|---|---|---|---|---|
| Pitch (Std) | -.65 | .14 | -4.49 | .01 |
| AWL Sublist-3 Normalized | .39 | .16 | 2.48 | .01 |
| Root TTR FW | .65 | .15 | 4.34 | .01 |
| Kuperman AoA (CW) | .90 | .17 | 5.39 | .01 |
| Number of moving objects | .56 | .15 | 3.72 | .01 |
| Moving Duration Difference (mean) | .81 | .15 | 5.46 | .01 |

*Note.* Constant term= 16.95; $\beta$ is the standardized beta coefficient; SE is standard error

The fitted video-level regression model included 7 multimodal complexity features; a single acoustic feature (*Std pitch*), three lexical features (*AWL Sublist_3 normalized, Root TTR FW*, and *Kuperman AoA CW*), and two visual features (*number of moving objects* and m*average of moving duration difference)*). These features jointly explained 44% of the variance in the video difficulty ratings. The correlation between the truth values and predicted values was $r$= 66. Overall, the model is performing reasonably well compared with the performance of the video-level model developed in Study 3 (Chapter 8) using the full set of multimodal complexity features.

#### 9.2.3.2 Stage Two: Estimating Video Segment Difficulty

The estimated difficulty of each video segment in the video corpus was derived using the coefficients and regression term in model one (Step 1). This process was implemented into steps. First, for each video segment of length 5 seconds, we extracted and computed the multimodal complexity features used in model one from the videotexts in the SLVC corpus. It should be noted that while computing multimodal complexity at segment and video level both involve averaging values to derive single values, the extraction of complexity features at the segment level was more involving and required further preprocessing. Specifically, the orthographic transcriptions should be aligned to the audio tracks to generate a phone-level segmentation for each word. For this purpose, I used the Gentle aligner—a python tool [1] that was built on top of Kaldi force alignment engine[2]—for identifying which time segments in the speech data corresponding to particular words in the transcription data. This alignment was necessary to compute lexical features for each video segment.

After extracting and computing all multimodal complexity features, this resultant matrix is of size (7, 34,363). Using model one, we estimate the difficulty for each video segment. Table 9.2 shows basic descriptive statistics of the new corpus.

Table 9.2: Descriptive Statistics for Level-One Model Multimodal Complexity Features

| Features | Mean | STD | Min | Max |
|---|---|---|---|---|
| Pitch std | 79.06 | 40.54 | 0 | 320.82 |
| TTR FW | .81 | .23 | 0 | 1 |
| AWL Sub-list 3 AW | .01 | .03 | 0 | 1 |
| AoA CW | 32.76 | 16.59 | 0 | 353.46 |
| Number of moving object | 6.81 | 3.99 | 0 | 75 |
| Moving object Duration Mean | .23 | .33 | 0 | 5.50 |

The average of all video segment difficulty is 16.95 and the standard deviation is 1.95. The minimum and maximum values are 9 and 32, respectively. Figure 9.4 shows the shape of the distributions, central values, and variability for randomly selected videos.

[1] https://github.com/lowerquality/gentle
[2] https://github.com/kaldi-asr/kaldi

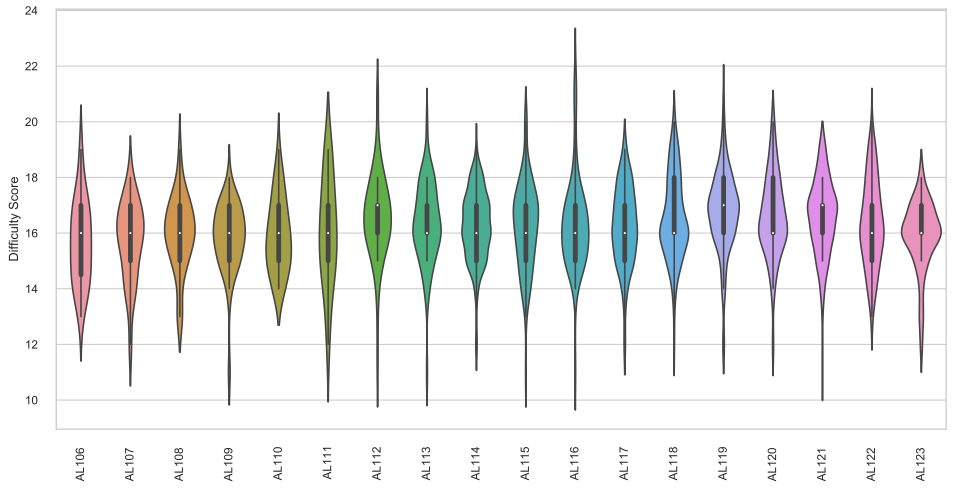



FIGURE 9.4: The Distributions, Central Values, and Variability of Randomly Selected Videos

#### 9.2.3.3 Stage Three: Developing a Time Series Forecasting Model

Despite that LSTM RNN–and artificial neural networks in general–are capable of modeling complex patterns in data without having to manually extract carefully designed features or perform feature engineering, I only considered a relatively small number of multimodal complexity features (n= 6) for training the model. This is mainly because extracting and computing multimodal complexity features is a very laborious and computationally expensive process.

### 9.2.4 Evaluating Model Forecasting

Because our dataset is substantially large (30,363 video segments), in this study, I used the out-of-sample approach for evaluating the performance of the models in accurately forecasting the difficulty of a future video segment. Specifically, I split the data into three folds, a training set, a validation set, and a testing set, using the 60-20-20 split method. The dataset was partitioned without shuffling to preserve the temporal order of the observations. The training and validation sets were used to train and tune the hyperparameters of the LSTM and GRU models. The hold-out test set was used only to evaluate the two models forecasting performance and generalizability on unseen video segments. For evaluating and comparing the performance of the trained models, the RMSE was used as the primary evaluation metric. The results showed that the stacked bidirectional GRU model

surpasses the performance of the other models in forecasting video segment difficulty. The evaluation results are shown in Table 9.3.

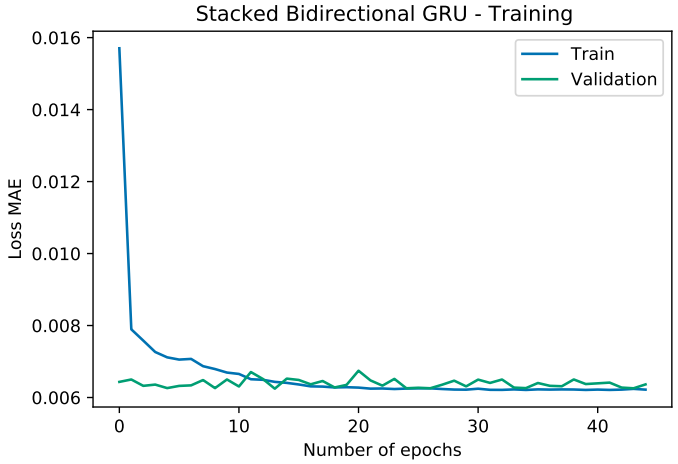



FIGURE 9.5: Stacked Bidirectional GRU - Training

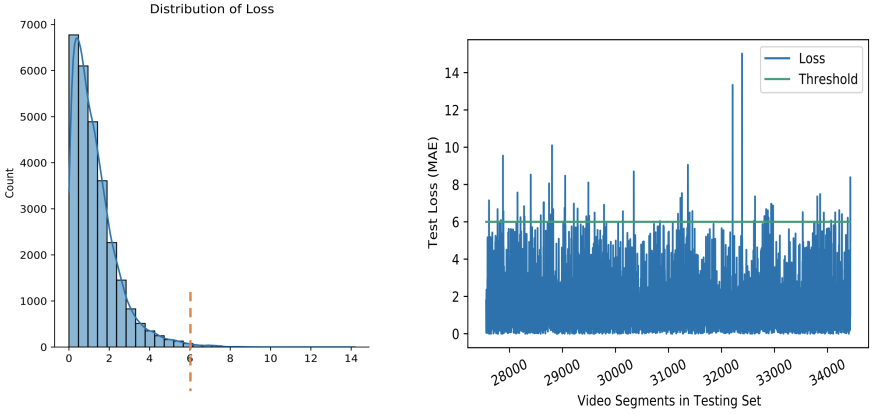



FIGURE 9.6: Manually Selecting Difficulty Threshold

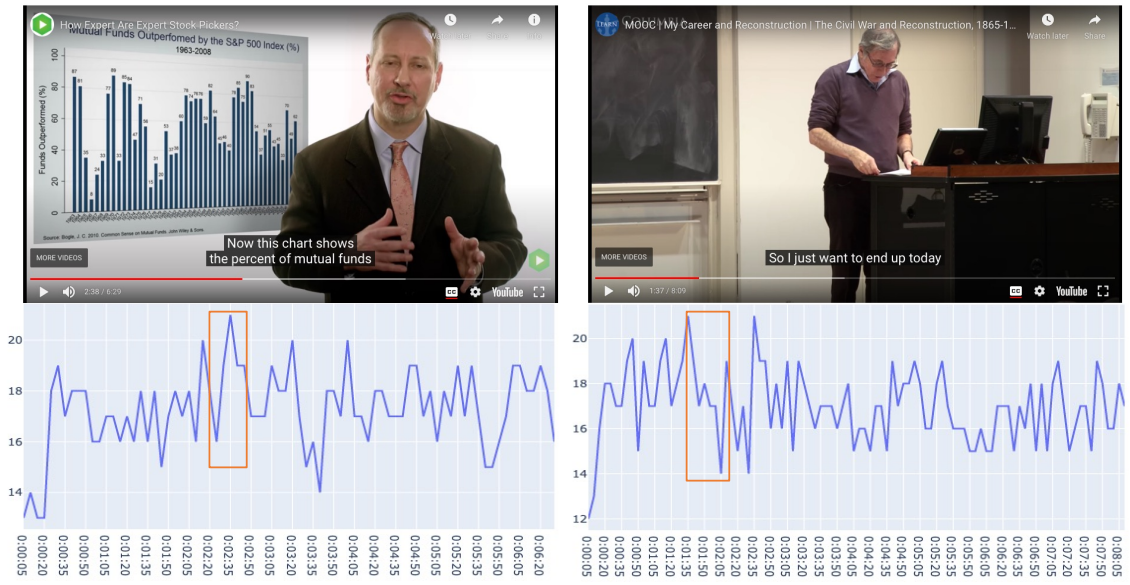
## 9.3 Qualitative Inspection of Model's Forecasting Ability

To inspect the forecasting ability of the trained model, I conducted qualitative analyses on a small sample of randomly selected videotexts (n= 10). Through visualizing the difficulty scores forecasted by the model and comparing them with what is shown and said in the videotexts, the results demonstrated the trained model seems to give reliable estimates

TABLE 9.3: A Comparison Between the Model Performance on Forecasting Video Segment Difficulty

| Model | MAE Train | MSE Train | MAE Test | MSE Test |
|---|---|---|---|---|
| LSTM | 1.3 | 3.02 | 1.45 | 3.95 |
| GRU | 1.31 | 3.03 | 1.46 | 3.97 |
| Stacked LSTM | 1.31 | 3.03 | 1.46 | 3.97 |
| Stacked GRU | 1.3 | 3.01 | 1.46 | **3.94** |
| Bidirectional LSTM | 1.3 | 3 | 1.47 | 3.97 |
| Bidirectional GRU | 1.3 | 3 | 1.46 | **3.94** |
| Stacked Bidirectional LSTM | 1.31 | 3.03 | 1.46 | 3.99 |
| Stacked Bidirectional GRU | 1.3 | 3 | 1.46 | **3.92** |

of video segment difficulty. Figure 9.7 shows the model's prediction of difficulty in two academic lectures.




(A) An Increase in Video Segment Difficulty (B) A Decrease in Video Segment Difficulty

FIGURE 9.7: Qualitative Inspection of Model's Forecasting Ability

In the video segment highlighted in (A), the trained model predicted a spike in difficulty where the lecturer began the segment saying, *"Now this chart shows the percent of mutual funds"* and while he was talking, the figure in the back was appearing. Because new verbal and visual information was presented concurrently, the model rightfully predicted an increase in difficulty. In (B), the model predicted a decrease in difficulty as the lecturer began the segment telling a story, *'So I just want to end up today by just telling you a little story ...'*. The trained model forecasted a decrease in difficulty as the lecturer started to use

more simple and common lexis compared to the previous segments in which the lecturer used more complex words.

## 9.4 Contributions and Limitations

This study makes several contributions to videotext difficulty assessment. Firstly, the study proposed an unsupervised approach to forecasting video segment difficulty in real-time. The approach does not require human raters to evaluate the difficulty of video segments. Instead, a well-performing multimodal model was used to provide estimations of the difficulty of video segments difficulty which were then used as training data for neural networks. The results of validation and quantitative analyses revealed that the proposed approach yielded reliable prediction of video segment difficulty. Secondly, the approach allows users to manually adjust the difficulty threshold. The trained model can be useful for L2 researchers, video content designers, and assessment specialists who are interested in moment-to-moment forecasting of video difficulty.

These contributions notwithstanding, the study has several limitations that need to be acknowledged. First and foremost, the proposed approach was not validated with language users. Although qualitative analyses showed the model was able to forecast difficulty rightfully, it would be helpful if the model's forecasting ability is validated with actual language learners before any use of the model for research or pedagogical practices. Another limitation is that I only extracted a small set of multimodal complexity for modeling video segment difficulty because of the constraints of the length (window size) of video segments as well as the computational cost involved. Though the developed model had reasonable predictive power, it might not be the most predictive model. The assessment of video segment difficulty is strictly contingent upon the accuracy of the video-level model that generates them. In other words, the estimated difficulty of video segments is as accurate as of the model that generates them.

Furthermore, in this study, I did not develop a genre-based model for estimating and forecasting the difficulty in AL and GA videotext. The findings of Study 1 and Study 2 showed that developing a video genre-based model leads to better prediction of difficulty. Future studies should consider investigating the impact of video genre on real-time forecasting of video difficulty.

Other techniques can be used for time series forecasting which we did not explore in this study. Feature study should consider the applicability of traditional statistical models on the task of forecasting video segment difficulty. What is more, in this study we were mainly interested in forecasting video segment difficulty, describing and accounting for an underlying trend, linking explanatory variables to the criterion of interest, and assessing the impact of critical events.

## 9.5 Conclusion

In this chapter, I proposed an analytical approach for forecasting the difficulty of short video segments that relies on training neural networks for detecting abnormal increases or decreases in difficulty. The approach was validated quantitatively on a large corpus of short video segments (n=34, 363). The results showed the Stacked Bidirectional GRU trained model forecasted video segment difficulty reasonably well. A qualitative inspection of the model forecasting showed that the model correctly identified difficult segments in videos. Therefore, the trained model can be used as an inspection tool in areas such as video content development, language learning, and assessment.

# Chapter 10

# Conclusion

In this chapter, I first summarize the findings of the four studies enclosed in this thesis. Then I outline the overall contributions and limitations of the thesis. Finally, I propose short-termed and long-termed research agendas for future research in automated videotext difficulty assessment.

## 10.1 Introduction

The overarching objective of this thesis was to investigate what makes videotext difficult for language learners (at the CEFR B1 level), and by extension, how to accurately predict the learners' evaluation of the difficulty of two video genres, academic lectures, and government advertisements. In Chapter 2, I reviewed the literature on computational text difficulty assessment and outlined four major approaches to text difficulty assessment, namely traditional readability formulas, cognitively inspired approaches, machine learning-based approaches, and deep learning or neural network approaches. In general, the four computational approaches differ on at least two aspects: feature representation and algorithm complexity; with readability formulas being simple and interpretable and deep learning models being more accurate but less interpretable. Based on reviewing the literature, I proposed 6 principles for a more accurate assessment of text difficulty. I summarize these principles below:

1. **Defining text difficulty**: Given the polysomy of the term, a well-articulated definition of videotext difficulty not only can eliminate confusion but also helps researchers to select more appropriate measures for estimating difficulty.

2. **Choosing more appropriate measure/s for the criterion variable**: Thus defined, researchers need to select measures that best gauge their construct of text difficulty.

3. **An inclusive treatment of text complexity**: For modeling text difficulty as a function of textual complexity, researchers need to account for all possible sources of complexity. This can be, of course, very challenging, but grounding the development of text difficulty assessment on a theory or theories of text comprehension can help researchers where to look.

4. **Controlling for text variation**: If not controlled, text variation can result in an unreliable estimation of complexity in texts. For instance, indices of lexical diversity (e.g., TTR) are more sensitive to text length. Also, text genre has an impact on the accuracy of text difficulty measures.

5. **Controlling for learner factors**: Besides textual complexity, text difficulty is affected by learner factors, e.g., age, language proficiency. One way to attenuate the interference of learner factors is through the development of text difficulty assessment tools for a homogeneous group of language learners.

6. **Validation**: To provide reliable feedback on text difficulty, automated difficulty tools should be validated.

These principles were used for the development of computational measures of videotext difficulty. In Chapter 3, I defined videotext difficulty as an increase in the attentional and cognitive resources required for successful understanding of videotext. I also made a clear distention between videotext difficulty and other similar constructs, videotext complexity and comprehension. In Chapter 5, I introduced the Second Language Videotext Complexity (SLVC) corpus to be used in investigating and modeling videotext difficulty in two video genres, Academic Lectures, and Government Advertisement. The corpus contains 640 videotexts, 320 videotexts from each video genre. The videotexts were curated from the Internet using a priory defined selection criteria. The entire corpus was then rated for difficulty by 322 Saudi EFL language learners using a modified version of a difficulty rating scale developed by S. Yoon et al. (2016) originally for spoken text difficulty. The modified scale was tried out on a small sample of language learners before running

the rating sessions and the results showed that the scale was reliable. To control for language proficiency effect on difficulty ratings, all the recruited participates were at the B1 level of the CFER framework. The participants' language proficiency was determined by the Cambridge Language Proficiency test which all participants took a week before the study as a mandatory language placement test for their English intensive program at a large public university in Saudi Arabia.

## 10.2 Summary of Key Findings

### 10.2.1 Verbal Complexity and Videotext Difficulty

The primary objective of Study 1 (Chapter 7) was to investigate whether verbal complexity can explain and predict videotext difficulty. Verbal complexity is a multidimensional construct, and our review of previous studies has identified multifarious potential predictors of text difficulty (Chapter 3). Using various computational linguistic tools and techniques, I extracted and computed a wide range of verbal complexity features extracted from the audio streams and the spoken transcripts of the videos. The extracted verbal complexity features were then examined for their explanatory and predictive ability in a series of iterative analyses. The key findings of this study are discussed below.

#### 10.2.1.1 Verbal Complexity and Videotext Difficulty Assessment

The overall findings of Study 1 suggest that verbal complexity can be used to explain and predict difficulty in L2 videotexts. Specifically, a model trained using all types of verbal complexity features outperformed explained 37% of the variance of language learners' judgment of difficulty of videotexts in the SLVC corpus (both AL and GA videotexts). The model also did better in predicting videotext difficulty (RMSE: 2.75 vs. 2.81) than did models developed separately using acoustic, lexical, syntactic, and discourse complexity features.

#### 10.2.1.2 Contributions of acoustic and phonological, lexical, syntactic, and discourse complexity to videotext difficulty

Study 1 compared the prediction accuracy of different verbal complexity models developed independently using acoustic, lexical, syntactic, and discourse complexity features. The results showed that a regression model developed using only lexical complexity features had higher predictive accuracy than models developed separately using acoustic, syntactic, and discourse complexity features. This finding suggests that the lexical complexity and knowledge of vocabulary, in general, play an important role in videotext understanding. Next in prediction accuracy comes a model developed using acoustic and phonological complexity features. Taken together, the findings suggest that acoustic and lexical complexity negatively impact videotext understanding. Whereas issues related to syntactic and discourse complexity may not significantly contribute to language learners' perception of videotext difficulty at the least for the genres included in this study.

### 10.2.2 Visual Complexity and Videotext Difficulty

Unlike verbal complexity, little research has been conducted on visual complexity and its impact on L2 videotext understanding. To fill this void in the literature, in Chapter 7 I proposed several potential spatial and spatiotemporal complexity features for explaining and predicting language learners' perception of videotext difficulty. These features were motivated by key principles and theories in the visual perception and cognition literature. To streamline the process of extraction, computing, and visualizing the proposed complexity features, I developed AUVANA, a cross-platform desktop application. AUVANA leverages state-of-the-art techniques from the fields of computer vision and video/image processing.

Using AUVANA, I extracted and computed 78 visual complexity features. To explore the relationships between and among these features and whether they characterize some latent structures, a principal component analysis was performed on the newly proposed features. The results showed that the visual complexity features cluster onto 11 different components or factors. This finding indicates that the proposed visual complexity features tap into different dimensions of visual complexity in videotexts and that video visual complexity is more likely a multifaceted and multidimensional construct. Further research, of course, is needed to explore each of these dimensions in much more detail.

As far as the relationship between visual complexity and videotext difficulty is concerned, the results of Study 2 (Chapter 7) clearly showed that visual complexity contributes substantially to videotext difficulty. Specifically, the best visual complexity model explained 29% of the variance in videotext difficulty. The model also performed well on predicting the difficulty on unseen videotext (r= .51, RMSE = 2.90).

#### 10.2.2.1 Video Genre and Videotext Difficulty Assessment

The impact of video genre on the predictive accuracy of verbal and visual complexity models was investigated in Study 1 and Study 2, respectively. The results of both studies demonstrated that difficulty in AL and GA videotexts was best predicted using different sets of verbal and visual complexity features and hence lend further support to the impact of the videotext genre on the accuracy of difficulty prediction models (Sheehan et al., 2013, 2014). For example, indices related to lexical sophistication and diversity were found to be strong predictors for the difficulty in AL videotexts, whereas acoustical and syntactic complexity were good indicators of difficulty in GA videotexts. On the other hand, the frequency of visual texts per shot was associated with more difficult AL videotexts, whereas the frequency of faces was associated with less difficult GA videotexts.

### 10.2.3 Verbal Complexity vs Visual Complexity: What Modality Predicts Videotext Difficulty Better?

One of the questions addressed in this thesis was what modality, verbal or visual, better explain and predict language learners' perception of difficulty in AL and GA videotexts. The findings showed that both verbal and visual complexity equally contributes to videotext difficulty. Specifically, the best verbal complexity model explained 33% of the variance in learners' rating of difficulty for all videotexts and had an RMSE of 2.75. Whereas the best visual complexity explained 29% of the variance in difficulty in videotexts and had an RMSE of 2.90. This result points to the importance of incorporating both verbal and visual complexity into measures of difficulty in L2 videotexts.

### 10.2.4 Unimodal vs. Multimodal Videotext Models

While Study 1 and Study 2 investigated independently the relative contributions of verbal and visual complexity, respectively, Study 3 examined whether combining both verbal

and visual complexity into a single predictive model leads to better prediction accuracy than developing a unimodal predictive model developed using either modality. To investigate this question, I developed several ensemble machine learning models and compared their performances in predicting videotext difficulty. Overall, the results demonstrated that incorporating multimodal complexity models lead to better prediction accuracy than using either vocal, verbal, or visual complexity alone. Secondly, ensemble machine learning models outperformed simple regression multimodal models in predicting difficulty in videotexts.

### 10.2.5 Forecasting the Difficulty of Short Video Segments

The main aim of Study 4 was concerned with forecasting the difficulty of videotext in real-time. To this end, a novel approach was proposed to give an estimate for video segment (five seconds) difficulty in real-time. The approach was implemented in three stages. In the first stage, a well-performant multimodal regression model was developed and used to give an estimate of video segment difficulty (which then became the ground truth for the succeeding model). In the second stage, artificial neural networks were trained to forecast the estimated difficulty and their predictive powers were compared. In the final stage, the distribution of the training error of the best performing neural network model was used to select a threshold for determining difficulty in videotexts. Through inspecting a plot of the distribution errors, an optimal threshold value was manually chosen to determine video segment difficulty in the unseen video segment dataset.

## 10.3 Contributions and Implications

The findings of this research make several contributions to the field of applied linguistics. They are summed into three categories—theory and research, language pedagogy, and assessment—depending on where they contribute the most.

### 10.3.1 Contributions to Theory and Research

This research program was the first investigation into the impact of videotext complexity on the perception of videotext difficulty. The findings of this thesis provide invaluable

insights on how different dimensions of videotext complexity contribute to videotext difficulty and in turn to the degradation of L2 videotext comprehension. In particular, the results demonstrated that both verbal and visual complexities contribute to L2 learners' perception of videotext difficulty. This finding emphasizes the important role that visual imagery plays in L2 videotext understanding. Further, our analysis of visual complexity pointed to the multidimensional and multifaceted nature of video visual complexity. In particular, the construct encompasses issues related to visual saliency, clutter, motion, and the presence of human faces.

To build a comprehensive view, the thesis reviewed and synthesized empirical evidence for the role of verbal and visual complexity from numerous fields and subfields, including but not limited to, L2 listening, text/discourse comprehension, visual perception, and cognition. In the lack of a defined body of literature around videotext difficulty and complexity, the review conducted herein can be an initial point of reference for researchers who are interested in understanding the different sources of complexity and how they may impact videotext comprehension.

As far as the implications of this thesis to research are concerned, scientific investigations of videotexts benefit from automated measures of videotext difficulty. For example, using the computational models developed in this thesis, researchers can control for the impact of videotext difficulty in their video research by choosing videotexts that are within the ability of their participants. As the results of this thesis showed verbal and visual complexity have a detrimental effect on language learners' perceptions of the difficulty of a video lecture. Therefore, researchers—who may design experimental tasks falsely assuming the universality of visual imagery, and/or believing that it demands no special competencies from language learners, to begin with—should control for or attenuate the impact of video complexity in their studies.

Regarding video visual complexity, the thesis contributes to the field by proposing new measures of visual complexity in videotexts, several of which have shown to be strong predictors of visual complexity. Also, the thesis developed the automated video analysis (AUVANA) software which implements state-of-the-art computer vision and video processing algorithms to automatically, extract, compute, and visualize visual complexity in videotexts. The software was made freely available to the research community with the attention of building a community of researchers around the use of AI in video complexity research as well as helping researchers with no coding background or expertise in computer vision to easily extract and analyze visual complexity in their video data.

Another contribution that this thesis makes to videotext difficulty research is the construction and release of the SLVC corpus. The corpus is the first annotated corpus for use in video complexity and difficulty research. It contains 640 annotated videotexts in two video genres, academic lectures, and government advertisements. The corpus can be used to devise new measures of multimodal complexity, develop computational models, and benchmarking computational models of videotext difficulty.

Finally, researchers who are interested in investigating real-time processing and understanding of videotext may find the prediction of neural network-based models developed in Chapter 9 useful for their research, though the accuracy of the models needs to be examined thoroughly.

### 10.3.2 Contributions to Pedagogy and Material Design

The thesis makes several recommendations to multimodal language learning and teaching and video material designs. Given that both verbal and visual complexity impact language learner's understanding of videotext, language teachers and video content developers should pay more attention to the complexity in both modalities in order to select videotexts that are more accessible for language learners. We found that L2 learners at the B1 level experience less difficulty viewing L2 videos if spoken words are articulated clearly and when there is a high pitch variation in the lecturer's speech. We believe that variation in the speaker's pitch helps L2 learners recognize words from speech. Therefore, when producing video materials for language learners, we recommend that the lecturer or narrator should speak clearly, with proper intonation and articulation.

Regarding the use of visual imagery in videotexts, some insights can be gleaned from the findings of Study 2. For example, video lecture difficulty is strongly associated with the increase of video shots, visual texts per shot, and the number of moving objects in videotext. This finding is expected because language learners need to decode more information. Therefore, it is recommended that video material designers should avoid excessive use of visual texts and visual objects, more specifically within shots.

## 10.4 Limitations and Directions for Future Studies

The thesis has some limitations which can guide future research. These limitations are related to the extraction and computation of complexity features, the construction and

annotation of the SLVC corpus, and the applications and generalizability.

### 10.4.1 Complexity Features

In modeling videotext difficulty, I only focused on features that can be easily extracted and computed from the verbal and visual modalities. The explanatory purpose of this thesis may justify this approach. However, future research should consider developing measures that account for the difficulty that arises from the interaction between the different modalities in videotexts, inter alia, music, sounds, gestures, facial expressions, scientific symbols, and visual imagery. One way forward is to integrate insights from multimodality theories in the development of multimodal complexity features.

Another limitation is that I only looked at difficulty as determined by global indices of complexity. For instance, in calculating the speaking rate, I aggregated the local and averaged them to reflect the average rate of the speaker. However, speech rate is dynamic—and a speaker may speed up or slow down at different segments of the video, for example, to emphasize a point or explain a complex concept.

### 10.4.2 Corpus and Video Difficulty Rating

In this thesis, I only examined videotext complexity in two video genres, academic lectures, and government advertisements. It should be noted that in developing the SLVC corpus I only selected videotexts that do not have any copyright restrictions; therefore, the corpus may not be truly representative of all types of videos in each genre. Future studies may consider contribute to the development of the SLVC corpus or build and investigate complexity in other video genres (e.g., news, films, soap operas)

The video difficulty raring scale had some limitations that need to be acknowledged. Designed to not burden the study's participants, the rating scale did not address issues germane to visual complexity, learners' grammatical knowledge, and familiarity with the video topics. Arguably, these issues may impact learners' overall assessment of videotext difficulty and therefore should be investigated in future studies. Another area of research worth pursuing is the development and validation of new criterion measures; ideally, measures that are non-intrusive and have little or no interference with learners' watching experience.

### 10.4.3 Applications and Generalizability

The findings of this thesis should be viewed as a confirmation of the utility of computational measures in assessing and predicting L2 videotext difficulty. Given this thesis was exploratory, future studies should investigate how to improve the accuracy of automated measures using more sophisticated machine learning algorithms. Another limitation of this study was the videotext genre and the demographic, gender, and language proficiency of the participants. Further research and development of videotext difficulty assessment tools are needed to cater to different video genres and populations of language learners. Another research area of high potential and wide applications in language teaching, learning, research, and assessment is the development of automated systems of videotext difficulty that correspond to CEFR levels.

## 10.5 A Glimpse into Future

Besides addressing and improving upon the shortcomings outlined above, there are several applications and areas of research in automated videotext difficulty that I believe are worthy of pursuing. I group them into two categories, short and long-term agendas, based on my expectation of their occurrence in the future.

### 10.5.1 Short-term Agenda

1. Augmenting automated videotext difficulty systems with intrusive sensor technologies, e.g., eye-tracker, thermal camera, and/or electroencephalogram (EEG) data, to accurately gauge videotext difficulty in real-time (e.g., Andreessen, Gerjets, Meurers, & Zander, 2020).

2. Personalized video-based learning systems that provide different help options and feedback to language learners based on a judgment of the videotext difficulty, task, and learner's language learning experience.

3. Intelligent video-based language assessment systems that generate comprehension questions with varying levels of difficulty.

4. Developing video corpus analysis tools that leverage more sophisticated NLP and computer vision technologies to analyze video data at a large scale.

### 10.5.2 Long-term Agenda

1. Automatic videotext generation or simplification to make videos more accessible to language learners and people with cognitive or linguistic impairments.

2. An AI agent that mimics language learner's video comprehension processes to help us understand multimodal language learning and its intricacies.

## 10.6 Concluding Remarks

In conclusion, this thesis aimed at developing automated models for explaining and predicting videotext difficulty for use in language teaching, learning, assessment, and research. Overall, the results of this thesis underscore the role of multimodal complexity for predicting videotext difficulty and point to the utility of computational complexity measures for assessing videotext difficulty, akin to those already available for estimating difficulty in written and spoken language materials. I hope that the outcomes of this study serve to open a discussion on videotext complexity and difficulty and spark a wave of empirical investigations in this less studied yet more important area of research.

**A**

# APPENDIX

# A.1 Plain Language Statement in Arabic


مدرسة اللغات واللغويات

**المشروع:** دراسة ما يجعل الفيديوهات صعبة لمتعلمي اللغة الثانية

د. بول جروبا (باحث مسؤول)
ايميل: p.gruba@unimelb.edu.au تلفون: 897-8344-3-61 +

عماد الغامدي (طالب دكتوراه)
ايميل: eaalghamdi@student.unimelb.edu.au

**المقدمة**
نشكرك على اهتمامك بالمشاركة في هذا المشروع البحثي. ستزودك الصفحات القليلة التالية بمزيد من المعلومات حول مشاركتك التطوعية في هذا البحث. إذا كنت لا ترغب في المشاركة، فلا داعي لذلك. ويمكنك أيضًا الانسحاب بعد في أي وقت اثناء مشاركتك في هذا البحث.

**ما هو هذا البحث عنه؟**
يهدف هذا البحث إلى دراسة صعوبات التي يواجها متعلمي اللغة الإنجليزية عند مشاهدة فيديوهات باللغة الإنجليزية.

**ماذا سأفعل؟**
في حالة موافقتك على المشاركة، سيُطلب منك مشاهدة 8-10 مقاطع فيديو والإجابة على 5 أسئلة لكل مقطع فيديو. مقاطع الفيديو عبارة عن محاضرات أكاديمية قصيرة ومقاطع إعلانات حكومية.

**ما هي الفوائد المحتملة؟**
يهدف المشروع إلى مساعدة معلمي اللغة على فهم ما يجعل الفيديوهات صعبة لمتعلمي اللغة. قد يساعد ذلك المعلمين، على سبيل المثال، في توفير مواد تعليمية أفضل لطلابهم.

**ما هي المخاطر المحتملة؟**
لا توجد مخاطر.

**هل يجب على المشاركة؟**
لا. المشاركة طوعية تمامًا. أنت قادر على الانسحاب (الخروج) في أي وقت.

**هل سأسمع عن نتائج هذا المشروع؟**
نعم يمكنك الاتصال بي للحصول على مزيد من المعلومات حول النتائج.

**ماذا سيحدث للمعلومات عني؟**
أنا أخذ سرية معلوماتك على محمل الجد. سيتم استبدال اسمك ومعلوماتك الشخصية الأخرى برموز عشوائية. بالإضافة إلى ذلك، لن يتم الكشف عن هويتك واجاباتك وغيرها من المعلومات لأي شخص آخر.

**أين يمكنني الحصول على مزيد من المعلومات؟**
إذا كنت ترغب في الحصول على مزيد من المعلومات حول المشروع، يرجى الاتصال بالباحثين:
PROF Paul Gruba (p.gruba@unimelb.edu.au) / Aأو عماد الغامدي
ealghamdi@student.unimelb.edu.au).)

## A.2 Difficulty Rating Scale in Arabic

1. كيف تقيم فهمك للفيديو؟

a) **اقل من ٦٠٪**
b) ٧٠٪
c) ٨٠٪
d) ٩٠٪
e) ١٠٠٪

2. ما مقدار المعلومات التي يمكنك تذكرها في الفيديو؟

a) **اقل من ٦٠٪**
b) ٧٠٪
c) ٨٠٪
d) ٩٠٪
e) ١٠٠٪

3. عدد الكلمات التي فاتتك أو لم تفهمها

a) أكثر من 10 كلمات
b) 6-10كلمات
c) 3-5كلمات
d) 1-2كلمات
e) لا شيء

4. كانت سرعة الكلام في الفيديو

a) سريع
b) سريع إلى حد ما
c) ليس سريع ولا بطيء
d) بطيء بعض الشيء
e) بطيء

5. اعتقد بان الفيديو

a) اعلى من مستواي جداً
b) اعلى من مستواي
c) مناسب لمستواي
d) اقل من مستواي
e) اقل من مستواي جداً

# A.3 Consent Form in Arabic

THE UNIVERSITY OF MELBOURNE

## نموذج الموافقة

جامعة ملبورن
مدرسة اللغات واللغات

المشروع: دراسة صعوبات الفيديوهات لمتعلمي اللغة الإنجليزية

باحث مسؤول: د. بول قروبا
باحث (طالب): عماد احمد الغامدي

اسم المشارك:

1. اتعهد بان عمري 18 او اكبر.
2. أوافق على المشاركة في هذا المشروع، وقد تم شرح التفاصيل الخاصة بالمشروع، وقد تم توفير بيان مكتوب بلغة واضحة لي لاحتفظ به.
3. أتفهم أن الغرض من هذا البحث هو دراسة ما يجعل الفيديوهات صعبة لمتعلمي اللغة الانجليزية.
4. أدرك أن مشاركتي في هذا المشروع مخصصة لأغراض البحث فقط.
5. أقر بأنه قد تم شرح لي التأثيرات المحتملة من مشاركتي في هذا المشروع البحثي.
6. في هذا المشروع، سوف يُطلب مني مشاهدة بعض الفيديوهات والاجابة على بعض الأسئلة.
7. أدرك أن مشاركتي تطوعية وأنني حر في الانسحاب من هذا المشروع في أي وقت دون وقوع أي ضرر علي.
8. أدرك أنه سيتم الاحتفاظ بمعلوماتي التي قدمتها في هذا البحث في جامعة ملبورن وسيتم اتلافها بعد 5 سنوات.
9. لقد أبلغت بأن المعلومات الخاصة بي سيتم حمايتها وحفظها ولا يسمح لغير الباحثين الرئيسين بالاطلاع عليها.
10. أتفهم أنه بعد التوقيع على نموذج الموافقة هذا وإعادته، سيحتفظ به الباحث.

توقيع المشارك: _____ التاريخ: _____

## A.4 A Summary of L2 Video Studies from 2000 to 2020

Table A.1: A Summary of L2 Video Studies from 2000 to 2020

| Study | L2 | Skill | Category | How |
|---|---|---|---|---|
| Chung (2002) | English | Listening | Advance Organizers | External Sources (TV Serise) |
| Wilberschied and Berman (2004) | Chinese | Listening | Advance Organizers | External Sources (TV Shows) |
| Ambard and Ambard (2012) | Spanish | Listening | Advance Organizers | |
| Coniam (2001) | English | Listening | Assessment | |
| Ockey (2007) | English | Listening | Assessment | |
| Wagner (2007) | English | Listening | Assessment | |
| Wagner (2008) | English | Listening | Assessment | |
| Wagner (2010b) | English | Listening | Assessment | |
| Wagner (2010c) | English | Listening | Assessment | |
| Cubilo and Winke (2013) | English | Listening | Assessment | External Sources (Test) |
| Wagner (2013) | English | Listening | Assessment | |
| Batty (2014) | English | Listening | Assessment | |
| Koyama and Ockey (2016) | English | Listening | Assessment | |
| Lesnov (2018) | English | Listening | Assessment | |
| Batty (2020) | English | Listening | Assessment | |
| P. Markham (1999) | English | Listening | Captions & Subtitle | Caption Speed, Synatctic Complexity |
| Chung (1999) | English | Listening | Captions & Subtitle | External Sources (Textbook) |
| P. L. Markham, Peter, and McCarthy (2001) | Spanish | Listening | Captions & Subtitle | External Sources (Textbook) |
| Pujolă (2002) | English | Listening | Captions & Subtitle | |
| G. Taylor (2005) | Spanish | Listening | Captions & Subtitle | Pilot Study |
| Winke et al. (2010) | A, C, S, R | Listening | Captions & Subtitle | |
| Lwo and Lin (2012) | English | Vocabulary | Captions & Subtitle | |
| Winke, Gass, and Sydorenko (2013) | A, C, R, S | | Captions & Subtitle | Expert Evaluation |
| J. C. Yang and Chang (2014) | English | Listening | Captions & Subtitle | |

**Table A.1 continued from previous page**

| Study | L2 | Skill | Category | How Difficulty is Assessed |
|---|---|---|---|---|
| H.-Y. Yang (2014) | English | Listening | Captions & Subtitle | |
| Montero Perez et al. (2014) | French | Listening | Captions & Subtitle | Expert Evaluation |
| Mirzaei, Meshgi, Akita, and Kawahara (2017) | English | Listening | Captions & Subtitle | Speech Rate, Lexical Frequency |
| Rodgers and Webb (2017) | English | Listening | Captions & Subtitle | |
| Montero Perez, Peters, and Desmet (2018) | French | Vocabulary | Captions & Subtitle | Expert Evaluation |
| Mestres and Pellicer-Sánchez (2019) | English | Listening | Captions & Subtitle | Readability formulas |
| Teng (2019) | English | Listening | Captions & Subtitle | Speech Rate, Lexical Frequency, Expert Evaluation |
| Montero Perez (2019) | French | Vocabulary | Captions & Subtitle | Lexical Coverage |
| Suárez and Gesa (2019) | English | Vocabulary | Captions & Subtitle | Lexical Coverage |
| Pujadas and Muñoz (2019) | English | Vocabulary | Captions & Subtitle | Lexical Coverage |
| Vanderplank (2019) | F, G, I, S | | Captions & Subtitle | |
| Cintrón-Valentín, García-Amaya, and Ellis (2019) | Spanish | Grammar | Captions & Subtitle | Lexical Coverage |
| Winke et al. (2019) | Spanish | Listening | Captions & Subtitle | |
| Gass et al. (2019) | English | Listening | Captions & Subtitle | |
| Y. T. Wang (2019) | English | Listening | Captions & Subtitle | |
| Wisniewska and Mora (2020) | English | Listening | Captions & Subtitle | |
| M. Lee and Révész (2020) | English | Grammar | Captions & Subtitle | |
| Pattemore and Muñoz (2020) | English | Grammar | Captions & Subtitle | |
| Hsieh (2020) | English | Listening | Captions & Subtitle | Pilot Study, Lexical and Syntactic complexity |
| Weyers (1999) | Spanish | Speaking | Language Skill | |
| Herron, York, Corrie, and Cole (2006) | French | Listening | Language Skill | |
| Dahl and Ludvigsen (2014) | English | Listening | Language Skill | Material Design |
| Arndt and Woore (2018) | English | Vocabulary | Language Skill | Lexical Coverage |
| Peters (2019) | English | Vocabulary | Language Skill | Lexical Coverage |
| Y. Feng and Webb (2019) | English | Vocabulary | Language Skill | Lexical Coverage |

**Table A.1 continued from previous page**

| Study | L2 | Skill | Category | How Difficulty is Assessed |
|---|---|---|---|---|
| K. M. Wong and Samudra (2019) | English | Vocabulary | Language Skill | External Sources, Lexical Frequency |
| Ryan and Granville (2020) | English | Pragmatics | Language Skill | |
| Y. Feng and Webb (2020) | English | Vocabulary | Language Skill | |
| Pujadas and Muñoz (n.d.) | English | Listening | Language Skill | Lexical coverage |
| Puimège and Peters (2020) | English | Vocabulary | Language Skill | Lexical Coverage |
| Montero Perez (2020) | French | Vocabulary | Language Skill | |
| Aldukhayel (2019) | English | Listening | Perception | Lexical Coverage |
| Herron, Cole, Corrie, and Dubreil (1999) | French | Culture | Strategy Use & Instruction | |
| Herron, Dubreil, Cole, and Corrie (2000) | French | Culture | Strategy Use & Instruction | |
| Williams and Thorne (2000) | Welsh | | Strategy Use & Instruction | |
| Herron, Dubreil, Corrie, and Cole (2002) | French | Culture | Strategy Use & Instruction | |
| Smidt and Hegelheimer (2004) | English | Listening | Strategy Use & Instruction | |
| Seo (2005) | Japanese | Listening | Strategy Use & Instruction | Pilot Study, expert evulation |
| Guichon and McLornan (2008) | English | Listening | Strategy Use & Instruction | |
| Cross (2009) | English | Listening | Strategy Use & Instruction | |
| Becker and Sturm (2017) | French | Listening | Strategy Use & Instruction | |
| Bozorgian and Alamdari (2018) | English | Listening | Strategy Use & Instruction | External Sources (Test) |
| Cárdenas-Claros (2020) | English | Listening | Strategy Use & Instruction | Subjective rating |
| H. J. H. Chen (2011) | English | Listening | Tool | |
| C.-M. Chen, Li, and Lin (2020) | English | Listening | Tool | |
| Gruba (2004) | Japanese | Listening | Visual processing | |
| Sueyoshi and Hardison (2005) | English | Listening | Visual processing | |
| Gruba (2006) | Japanese | Listening | Visual processing | Pilot Study |
| Cross (2011) | English | Listening | Visual processing | |
| Suvorov (2014b) | English | Listening | Visual processing | |

**Table A.1 continued from previous page**

| Study | L2 | Skill | Category | How Difficulty is Assessed |
|---|---|---|---|---|
| Puimège and Peters (2019) | English | Vocabulary | Visual processing | Lexical Coverage |
| Rodgers and Webb (2020) | English | Vocabulary | Visual processing | |

## A.5 The Results of Inter-rater Reliability Tests on all Video Rating Groups

| Group | Raters(n) | ICC | F | df1 | df2 | *p* | Lower-band | Upper-band |
|---|---|---|---|---|---|---|---|---|
| 1 | 5 | .73 | 4.9 | 9 | 36 | 0 | .379 | .92 |
| 2 | 5 | .23 | 1.4 | 9 | 36 | .25 | -.59 | .76 |
| 3 | 4 | .47 | 2.4 | 9 | 27 | .035 | -.067 | .83 |
| 4 | 4 | .77 | 4.3 | 9 | 27 | 0 | .4 | .93 |
| 5 | 4 | .51 | 3 | 9 | 27 | .014 | .00145 | .84 |
| 6 | 5 | .22 | 1.93 | 9 | 36 | .078 | -.094 | .64 |
| 7 | 4 | .73 | 5.9 | 9 | 27 | 0 | .321 | .92 |
| 8 | 6 | .4 | 2.8 | 9 | 45 | .011 | .0078 | .78 |
| 9 | 5 | .93 | 20 | 9 | 36 | 0 | .81 | .98 |
| 10 | 4 | .71 | 7 | 9 | 27 | 0 | .248 | .92 |
| 11 | 5 | .46 | 4.7 | 9 | 36 | 0 | .054 | .81 |
| 12 | 3 | .63 | 3.9 | 9 | 18 | 0 | .072 | .89 |
| 13 | 4 | .72 | 6 | 9 | 27 | 0 | .272 | .92 |
| 14 | 5 | .53 | 3.5 | 9 | 36 | 0 | .087 | .85 |
| 15 | 4 | .52 | 3.8 | 9 | 27 | 0 | .0279 | .84 |
| 16 | 4 | .69 | 7.6 | 9 | 27 | 0 | .2013 | .91 |
| 17 | 3 | .36 | 1.7 | 9 | 18 | .15 | -.41 | .8 |
| 18 | 3 | .54 | 2.6 | 9 | 18 | .041 | -.096 | .87 |
| 19 | 4 | .6 | 4 | 9 | 27 | 0 | .109 | .87 |
| 20 | 4 | .8 | 8.5 | 9 | 27 | 0 | 0.45 | .94 |
| 21 | 5 | .6 | 3.9 | 9 | 36 | 0 | 0.165 | .87 |
| 22 | 3 | .61 | 4.3 | 9 | 18 | 0 | 0.017 | .88 |
| 23 | 3 | .81 | 6.2 | 9 | 18 | 0 | 0.46 | .95 |
| 24 | 3 | .64 | 3.6 | 9 | 18 | .01 | .089 | .9 |
| 25 | 3 | .84 | 6.3 | 9 | 18 | 0 | .55 | .96 |
| 26 | 4 | .63 | 2.8 | 9 | 27 | .02 | .068 | .9 |
| 27 | 4 | .63 | 3.5 | 9 | 27 | .01 | .148 | .89 |
| 28 | 3 | .48 | 2.2 | 9 | 18 | .07 | -.208 | .84 |
| 29 | 3 | .42 | 1.7 | 9 | 18 | .16 | -.67 | .84 |
| 30 | 3 | .56 | 2.3 | 9 | 18 | .06 | -.243 | .88 |
| 31 | 4 | .68 | 3.1 | 9 | 27 | .01 | .163 | .91 |
| 32 | 3 | .62 | 2.7 | 9 | 18 | .04 | -.052 | .9 |
| 33 | 3 | .55 | 2.3 | 9 | 18 | 0.07 | -.23 | .87 |
| 34 | 4 | .83 | 6 | 9 | 27 | 0 | .57 | .95 |
| 35 | 3 | .52 | 2.7 | 9 | 18 | .03 | -.086 | .85 |
| 36 | 7 | .63 | 3.1 | 9 | 54 | .01 | .199 | .89 |
| 37 | 6 | .86 | 9 | 9 | 45 | 0 | .68 | .96 |

**Table A.2 continued from previous page**

| Group | Raters(n) | ICC | F | df1 | df2 | p | Lower-band | Upper-band |
|---|---|---|---|---|---|---|---|---|
| 38 | 6 | .7 | 5.4 | 9 | 45 | 0 | .349 | .91 |
| 39 | 7 | .75 | 4.6 | 9 | 54 | 0 | .437 | .92 |
| 40 | 8 | .94 | 16 | 9 | 63 | 0 | .86 | .98 |
| 41 | 6 | .88 | 10.9 | 9 | 45 | 0 | .71 | .97 |
| 42 | 6 | .48 | 2.4 | 9 | 45 | .03 | -.00782 | .83 |
| 43 | 6 | .49 | 2.5 | 9 | 45 | .02 | .0048 | .84 |
| 44 | 6 | .89 | 16 | 9 | 45 | 0 | .72 | .97 |
| 45 | 7 | .71 | 4.1 | 9 | 54 | 0 | .368 | .91 |
| 46 | 6 | .85 | 7.1 | 9 | 45 | 0 | .65 | .96 |
| 47 | 6 | .74 | 4.5 | 9 | 45 | 0 | .417 | .92 |
| 48 | 7 | .64 | 3.7 | 9 | 54 | 0 | .253 | .89 |
| 49 | 7 | .63 | 2.9 | 9 | 54 | .01 | .182 | .89 |
| 50 | 9 | .74 | 8.2 | 9 | 72 | 0 | .445 | .92 |
| 51 | 6 | .62 | 2.8 | 9 | 45 | .01 | .137 | .89 |
| 52 | 7 | .56 | 2.5 | 9 | 54 | .02 | .0585 | .87 |
| 53 | 6 | .6 | 2.8 | 9 | 45 | .01 | .1176 | .88 |
| 54 | 6 | .77 | 6.2 | 9 | 45 | 0 | .465 | .93 |
| 55 | 5 | .2 | 1.3 | 9 | 36 | .28 | -.733 | .76 |
| 56 | 3 | .01 | 0 | 9 | 18 | .47 | -1.13 | .69 |
| 57 | 4 | .01 | 0 | 9 | 27 | .46 | -.58 | .06 |
| 58 | 3 | .08 | 1.12 | 9 | 18 | .40 | -1.02 | .72 |
| 59 | 3 | .15 | 1.22 | 9 | 18 | .34 | -.82 | .73 |
| 60 | 3 | .62 | 2.7 | 9 | 18 | .04 | *.52 | .9 |
| 61 | 2 | .11 | 1.36 | 9 | 9 | .33 | -.3 | .61 |
| 62 | 3 | .33 | 1.6 | 9 | 18 | .19 | -.51 | .79 |
| 63 | 5 | .09 | 1.1 | 9 | 36 | .39 | -1.27 | .74 |
| 64 | 6 | .64 | 2.8 | 9 | 45 | .011 | .149 | .9 |

## A.6 AUVANA Metrics

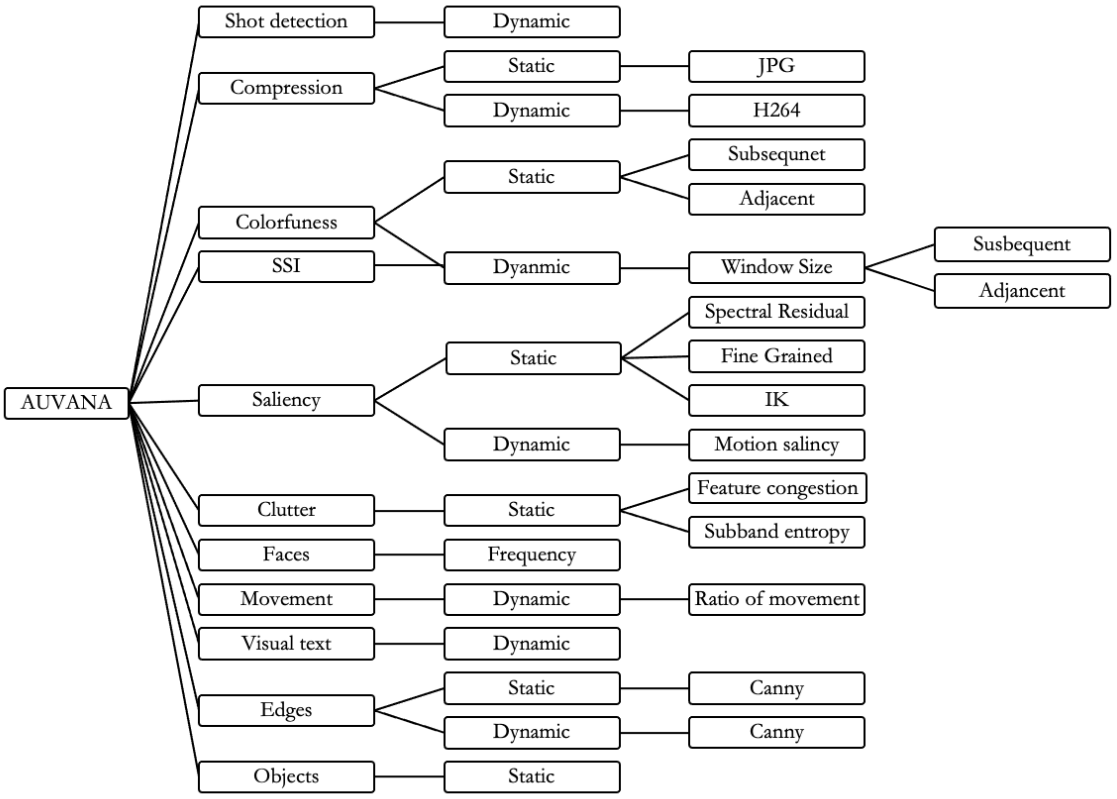



Figure A.1: A Diagram of AUVANA Metrics

## A.7 Component Correlation Matrix

| Component | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | - | | | | | | | | | | | |
| 2 | .13 | | | | | | | | | | | |
| 3 | -.04 | 0 | | | | | | | | | | |
| 4 | .00 | .18 | -.03 | | | | | | | | | |
| 5 | .26 | -.06 | .16 | .05 | | | | | | | | |
| 6 | -.21 | -.21 | -.10 | -.10 | -.17 | | | | | | | |
| 7 | .10 | .20 | -.08 | .17 | .01 | -.17 | | | | | | |
| 8 | -.15 | .12 | -.21 | .05 | -.12 | -.06 | .07 | | | | | |
| 9 | .10 | .17 | -.06 | .18 | .01 | -.12 | .03 | .15 | | | | |
| 10 | .21 | .24 | .08 | .08 | .23 | -.13 | .10 | .02 | .16 | | | |
| 11 | .18 | .16 | -.01 | -.03 | -.01 | .02 | .08 | -.05 | -.08 | .05 | | |
| 12 | -.05 | -.07 | .11 | -.05 | .02 | .15 | -.15 | -.06 | -.08 | 0 | 0 | - |

*Note.* Extraction Method: Principal Component Analysis
Rotation Method: Oblimin with Kaiser Normalization.

## A.8 Correlation and Descriptive Statistics for Verbal Complexity Features

| Feature | *r* | Sig. | *M* | *SD* |
|---|---|---|---|---|
| Number of syllables | .34 | .01 | 1033.75 | 932.75 |
| Number of pauses | .22 | .00 | 57.31 | 62.50 |
| Duration (s) | .34 | .00 | 268.03 | 218.49 |
| Phonation time(s) | .35 | .00 | 224.15 | 187.70 |
| Speech rate | .21 | .00 | 3.77 | 0.56 |
| Articulation rate | .12 | .00 | 4.47 | 0.41 |
| ASD | -.11 | .01 | 0.23 | 0.02 |
| Pitch (Min) | -.01 | .87 | 60.76 | 2.84 |
| Pitch (Max) | .07 | .07 | 586.56 | 43.41 |
| Pitch (Mean) | -.03 | .42 | 154.21 | 28.73 |
| Pitch (STD) | -.28 | .00 | 66.44 | 21.92 |
| k1 BNC-COCA | -.37 | .00 | 71.87 | 9.93 |

**Table A.3 continued from previous page**

| Feature | *r* | Sig. | *M* | *SD* |
|---|---|---|---|---|
| k2 BNC-COCA | .26 | .00 | 11.28 | 4.15 |
| k3 BNC-COCA | .40 | .00 | 6.65 | 3.73 |
| k1-2 BNC-COCA | -.37 | .00 | 83.15 | 7.03 |
| k1-3 BNC-COCA | -.22 | .00 | 89.80 | 4.98 |
| k1-4 BNC-COCA | -.18 | .00 | 91.53 | 4.41 |
| k1 5 BNC-COCA | -.18 | .00 | 91.53 | 4.41 |
| k1 6 BNC-COCA | -.17 | .00 | 92.12 | 4.18 |
| k1 7 BNC-COCA | -.17 | .00 | 92.52 | 4.00 |
| k1 8 BNC-COCA | -.18 | .00 | 92.85 | 3.87 |
| k1 BNC-COCA (CW) | -.05 | .22 | 45.58 | 5.76 |
| k2 BNC-COCA (CW) | .26 | .00 | 10.91 | 4.05 |
| k3 BNC-COCA (CW) | .39 | .00 | 6.50 | 3.64 |
| k1-2 BNC-COCA (CW) | .15 | .00 | 56.49 | 5.20 |
| k1-3 BNC-COCA (CW) | .34 | .00 | 62.99 | 6.40 |
| k1-4 BNC-COCA (CW) | .37 | .00 | 64.70 | 6.61 |
| Lexical sophistication (CW) | .03 | .38 | 0.23 | 3.37 |
| Lexical sophistication$_2(CW)$ | .23 | .00 | 0.13 | 0.05 |
| Lexical sophistication noun | .03 | .39 | 0.14 | 2.06 |
| Lexical sophistication verb | .03 | .38 | 0.04 | 0.63 |
| AWL (CW) | .30 | .00 | 7.49 | 3.53 |
| AWL (AW) | .30 | .00 | 7.68 | 3.56 |
| All AWL Normed | .27 | .00 | 0.05 | 0.03 |
| AWL Sublist 1 Normed | .17 | .00 | 0.01 | 0.01 |
| AWL Sublist 2 Normed | .18 | .00 | 0.01 | 0.01 |
| AWL Sublist 3 Normed | .21 | .00 | 0.00 | 0.01 |
| AWL Sublist 4 Normed | .13 | .00 | 0.01 | 0.01 |
| AWL Sublist 5 Normed | .12 | .00 | 0.01 | 0.01 |
| AWL Sublist 6 Normed | .06 | .11 | 0.00 | 0.01 |
| AWL Sublist 7 Normed | .11 | .00 | 0.00 | 0.00 |
| AWL Sublist 8 Normed | -.04 | .31 | 0.00 | 0.01 |
| AWL Sublist 9 Normed | .07 | .06 | 0.00 | 0.00 |
| AWL Sublist 10 Normed | .13 | .00 | 0.00 | 0.00 |
| All AFL Normed | -.18 | .00 | 0.05 | 0.03 |
| Core AFL Normed | -.04 | .33 | 0.02 | 0.02 |
| Spoken AFL Normed | -.24 | .00 | 0.03 | 0.02 |
| Written AFL Normed | .03 | .39 | 0.01 | 0.01 |
| ASWL level$_1$ | -.18 | .00 | 0.55 | 0.06 |
| ASWL level$_2$ | .03 | .40 | 0.03 | 0.02 |
| ASWL level$_3$ | .24 | .00 | 0.01 | 0.01 |
| ASWL level$_4$ | -.03 | .38 | 0.00 | 0.00 |
| Lexical density (type) | .40 | .00 | 0.64 | 0.08 |

**Table A.3 continued from previous page**

| Feature | *r* | Sig. | *M* | *SD* |
|---|---|---|---|---|
| Lexical density token | .11 | .01 | 0.44 | 0.05 |
| TTR (CW) | .02 | .63 | 0.62 | 0.14 |
| TTR (FW) | -.22 | .00 | 0.30 | 0.14 |
| Root TTR(AW) | .42 | .00 | 8.28 | 2.21 |
| Root TTR (CW) | .41 | .00 | 8.16 | 3.07 |
| Root TTR (FW) | .23 | .00 | 3.83 | 0.47 |
| Log TTR (CW) | .18 | .00 | 0.90 | 0.04 |
| Log TTR (FW) | -.17 | .00 | 0.76 | 0.05 |
| TTR (AW) | -.05 | .25 | 0.45 | 0.14 |
| Log TTR (AW) | .04 | .28 | 0.86 | 0.05 |
| Noun TTR | .03 | .48 | 0.61 | 0.16 |
| Verb TTR | .05 | .23 | 0.70 | 0.14 |
| Adj. TTR | -.04 | .29 | 0.70 | 0.17 |
| Root verb | .38 | .00 | 5.90 | 2.19 |
| Corrected verb | .38 | .00 | 4.17 | 1.55 |
| Verb variation | -.10 | .01 | 0.19 | 0.05 |
| Verb variation2 | -.10 | .01 | 0.19 | 0.05 |
| Noun variation | .03 | .40 | 0.26 | 0.08 |
| Adj. variation | .13 | .00 | 0.10 | 0.03 |
| Adv. variation | -.07 | .10 | 0.09 | 0.03 |
| MASS TTR (AW) | -.26 | .00 | 0.06 | 0.01 |
| MASS TTR (CW) | -.31 | .00 | 0.04 | 0.02 |
| MASS TTR (FW) | -.10 | .01 | 0.10 | 0.02 |
| MATTR50 (AW) | .22 | .00 | 0.73 | 0.05 |
| MATTR50 (CW) | .32 | .00 | 0.78 | 0.10 |
| MATTR50 (FW) | -.01 | .72 | 0.53 | 0.06 |
| MSTTR50 (AW) | .21 | .00 | 0.73 | 0.05 |
| MSTTR50 (CW) | .30 | .00 | 0.78 | 0.10 |
| MSTTR50 (FW) | -.01 | .81 | 0.52 | 0.07 |
| HDD42 (AW) | .26 | .00 | 0.78 | 0.11 |
| HDD42 (CW) | .27 | .00 | 0.77 | 0.24 |
| HDD42 (FW) | .17 | .00 | 0.54 | 0.14 |
| MTLD original (AW) | .19 | .00 | 53.92 | 18.36 |
| MTLD original (CW) | .31 | .00 | 87.59 | 63.17 |
| MTLD original (FW) | .05 | .19 | 18.63 | 4.85 |
| MTLD MA bi (AW) | .29 | .00 | 48.97 | 16.82 |
| MTLD MA bi (CW) | .27 | .00 | 60.11 | 66.60 |
| MTLD MA bi (FW) | .09 | .03 | 17.76 | 3.68 |
| MTLD MA wrap (AW) | .28 | .00 | 52.91 | 15.06 |
| MTLD MA wrap (CW) | .33 | .00 | 84.56 | 66.09 |
| MTLD MA wrap (FW) | .05 | .20 | 18.41 | 3.54 |

**Table A.3 continued from previous page**

| Feature | *r* | Sig. | *M* | *SD* |
|---|---|---|---|---|
| MRC Familiarity (AW) | -.21 | .00 | 593.51 | 4.47 |
| MRC Concreteness (AW) | -.15 | .00 | 298.81 | 13.23 |
| MRC Imageability (AW) | -.13 | .00 | 321.75 | 13.35 |
| MRC Meaningfulness (AW) | -.21 | .00 | 348.84 | 12.74 |
| MRC Familiarity (CW) | -.26 | .00 | 577.05 | 7.89 |
| MRC Concreteness (CW) | .03 | .45 | 350.09 | 20.72 |
| MRC Imageability (CW) | .04 | .35 | 378.16 | 20.34 |
| MRC Familiarity (FW) | -.05 | .23 | 610.26 | 2.47 |
| MRC Concreteness (FW) | -.23 | .00 | 249.36 | 14.28 |
| MRC Imageability (FW) | -.22 | .00 | 263.71 | 13.30 |
| MRC Meaningfulness (FW) | -.23 | .00 | 301.86 | 16.40 |
| Kuperman AoA (AW) | .37 | .00 | 5.40 | 0.39 |
| Kuperman AoA (CW) | .41 | .00 | 6.18 | 0.61 |
| Kuperman AoA (FW) | .07 | .08 | 4.38 | 0.14 |
| Brysbaert Concreteness Combined AW | -.16 | .00 | 2.48 | 0.13 |
| Brysbaert Concreteness Combined CW | -.06 | .15 | 2.83 | 0.17 |
| Brysbaert Concreteness Combined (FW) | -.20 | .00 | 2.05 | 0.14 |
| COCA spoken bi MI | .07 | .09 | 1.54 | 0.19 |
| COCA spoken bi MI2 | -.12 | .00 | 9.24 | 0.35 |
| COCA spoken bi T | .03 | .42 | 57.89 | 12.22 |
| COCA spoken bi DP | .06 | .13 | 0.06 | 0.01 |
| COCA spoken bi AC | .09 | .03 | 16277.82 | 4785.94 |
| COCA spoken tri MI | .03 | .42 | 2.57 | 0.29 |
| COCA spoken tri MI2 | .04 | .37 | 8.67 | 0.37 |
| COCA spoken tri T | .04 | .28 | 24.97 | 5.85 |
| COCA spoken tri DP | .05 | .24 | 0.01 | 0.00 |
| COCA spoken tri AC | .04 | .27 | 1889.39 | 1144.17 |
| COCA spoken tri 2 MI | .14 | .00 | 2.60 | 0.32 |
| COCA spoken tri 2 MI2 | .12 | .00 | 8.70 | 0.41 |
| COCA spoken tri 2 T | .06 | .16 | 24.99 | 5.61 |
| COCA spoken tri 2 DP | .09 | .02 | 0.15 | 0.03 |
| COCA spoken tri 2 AC | .05 | .25 | 1841.68 | 1099.44 |
| COCA academic bi MI | -.07 | .07 | 1.48 | 0.19 |
| COCA academic bi MI2 | -.01 | .72 | 8.64 | 0.28 |
| COCA academic bi T | .11 | .01 | 34.46 | 12.31 |
| COCA academic bi DP | .09 | .02 | 0.05 | 0.01 |
| COCA academic bi AC | .18 | .00 | 10913.53 | 4774.06 |
| COCA academic tri DP | .07 | .10 | 0.01 | 0.00 |
| COCA academic tri AC | .12 | .00 | 663.50 | 317.78 |
| COCA academic tri 2 MI | -.01 | .81 | 2.75 | 0.32 |
| COCA academic tri 2 MI2 | .06 | .13 | 8.07 | 0.38 |

**Table A.3 continued from previous page**

| Feature | *r* | Sig. | *M* | *SD* |
|---|---|---|---|---|
| COCA academic tri 2 T | .08 | .03 | 15.49 | 3.44 |
| COCA academic tri 2 DP | .10 | .01 | 0.16 | 0.04 |
| COCA academic tri 2 AC | .12 | .00 | 651.17 | 315.95 |
| Basic connectives | -.16 | .00 | 0.05 | 0.02 |
| Conjunctions | .02 | .63 | 0.03 | 0.01 |
| Disjunctions | -.09 | .03 | 0.00 | 0.01 |
| Lexical subordinators | -.10 | .01 | 0.02 | 0.01 |
| Coordinating conjuncts | -.25 | .00 | 0.01 | 0.01 |
| Addition | -.01 | .80 | 0.04 | 0.01 |
| Sentence linking | -.05 | .20 | 0.02 | 0.01 |
| Order | -.02 | .55 | 0.01 | 0.00 |
| Reason and purpose | .01 | .75 | 0.01 | 0.01 |
| All causal | -.21 | .00 | 0.02 | 0.01 |
| Positive causal | -.18 | .00 | 0.02 | 0.01 |
| Opposition | .06 | .11 | 0.01 | 0.01 |
| Determiners | -.03 | .51 | 0.10 | 0.02 |
| All demonstratives | -.17 | .00 | 0.03 | 0.01 |
| Attended demonstratives | -.07 | .06 | 0.01 | 0.01 |
| Unattended demonstratives | -.17 | .00 | 0.02 | 0.01 |
| All additive | .03 | .49 | 0.05 | 0.01 |
| All logical | -.04 | .31 | 0.04 | 0.01 |
| Positive logical | -.04 | .36 | 0.02 | 0.01 |
| Negative logical | .05 | .20 | 0.01 | 0.00 |
| All temporal | .02 | .65 | 0.01 | 0.01 |
| Positive intentional | -.11 | .01 | 0.01 | 0.01 |
| All positive | -.09 | .02 | 0.07 | 0.02 |
| All negative | -.02 | .54 | 0.01 | 0.01 |
| All connective | .02 | .59 | 0.07 | 0.02 |
| Pronoun density | -.12 | .00 | 0.05 | 0.02 |
| Pronoun noun ratio | -.18 | .00 | 0.25 | 0.14 |
| Repeated content lemmas | .09 | .02 | 0.30 | 0.08 |
| Repeated content and pronoun lemmas | .06 | .10 | 0.35 | 0.08 |
| Argument ttr | .03 | .42 | 0.46 | 0.14 |
| Bigram lemma ttr | .05 | .20 | 0.82 | 0.09 |
| Trigram lemma ttr | .20 | .00 | 0.94 | 0.05 |
| Adjacent overlap all sent | -.06 | .12 | 0.18 | 0.09 |
| Adjacent overlap all sent div seg | -.01 | .78 | 3.40 | 6.85 |
| Adjacent overlap binary all sent | .05 | .17 | 0.73 | 0.33 |
| Adjacent overlap cw sent | -.15 | .00 | 0.10 | 0.07 |
| Adjacent overlap cw sent div seg | -.03 | .49 | 1.16 | 3.85 |
| Adjacent overlap binary cw sent | .01 | .84 | 0.44 | 0.25 |

**Table A.3 continued from previous page**

| Feature | *r* | Sig. | *M* | *SD* |
|---|---|---|---|---|
| Adjacent overlap fw sent | .03 | .38 | 0.26 | 0.13 |
| Adjacent overlap fw sent div seg | .01 | .87 | 2.17 | 3.37 |
| Adjacent overlap binary fw sent | .08 | .05 | 0.68 | 0.32 |
| Adjacent overlap noun sent | -.14 | .00 | 0.11 | 0.08 |
| Adjacent overlap noun sent div seg | -.02 | .61 | 0.54 | 1.50 |
| Adjacent overlap binary noun sent | .03 | .51 | 0.29 | 0.22 |
| Adjacent overlap verb sent | -.08 | .04 | 0.10 | 0.08 |
| Adjacent overlap verb sent div seg | -.03 | .42 | 0.36 | 1.31 |
| Adjacent overlap binary verb sent | -.04 | .30 | 0.21 | 0.19 |
| Adjacent overlap adj sent | -.12 | .00 | 0.05 | 0.07 |
| Adjacent overlap adj sent div seg | -.03 | .41 | 0.09 | 0.41 |
| Adjacent overlap binary adj sent | -.02 | .67 | 0.07 | 0.11 |
| Adjacent overlap adv sent | -.10 | .02 | 0.07 | 0.08 |
| Adjacent overlap adv sent div seg | -.03 | .43 | 0.13 | 0.65 |
| Adjacent overlap binary adv sent | -.05 | .21 | 0.08 | 0.15 |
| Adjacent overlap pronoun sent | -.07 | .10 | 0.28 | 0.18 |
| Adjacent overlap pronoun sent div seg | -.07 | .07 | 0.48 | 0.68 |
| Adjacent overlap binary pronoun sent | -.07 | .06 | 0.34 | 0.26 |
| Adjacent overlap argument sent | -.17 | .00 | 0.15 | 0.09 |
| Adjacent overlap argument sent div seg | -.03 | .38 | 0.93 | 1.93 |
| Adjacent overlap binary argument sent | -.01 | .72 | 0.47 | 0.27 |
| Syn overlap sent noun | -.02 | .57 | 1.03 | 6.49 |
| Syn overlap sent verb | -.03 | .53 | 1.67 | 16.27 |
| MLS | .05 | .17 | 22.31 | 32.12 |
| MLT | .16 | .00 | 18.44 | 9.55 |
| MLC | .31 | .00 | 9.39 | 2.28 |
| C S | -.01 | .74 | 2.29 | 1.63 |
| VP T | .04 | .29 | 2.67 | 1.41 |
| C T | .02 | .58 | 1.98 | 1.03 |
| DC C | .16 | .00 | 0.41 | 0.11 |
| DC T | .04 | .38 | 0.87 | 0.73 |
| T S | -.08 | .05 | 1.13 | 0.28 |
| CT T | .11 | .00 | 0.47 | 0.17 |
| CP T | .12 | .00 | 0.40 | 0.34 |
| CP C | .14 | .00 | 0.20 | 0.15 |
| CN T | .20 | .00 | 2.09 | 1.37 |
| CN C | .37 | .00 | 1.05 | 0.37 |

## A.9 Correlation and Descriptive Statistics for Visual Complexity Features

| Feature | *r* | Sig. | *M* | *SD* |
|---|---|---|---|---|
| Frequency of Shots | 0.26 | 0.00 | 24.96 | 34.81 |
| Frequency of Visual Text | 0.10 | 0.01 | 64.01 | 129.49 |
| Frequency of Visual Text per Shot | -0.02 | 0.62 | 24.57 | 104.56 |
| Frequency of Visual Text per Second | -0.02 | 0.70 | 0.26 | 0.43 |
| Frequency of Visual Text per Minute | -0.02 | 0.70 | 15.70 | 25.83 |
| Frequency of speakers | -0.12 | 0.00 | 3.25 | 2.63 |
| Frequency of faces | 0.09 | 0.02 | 4.05 | 6.61 |
| Speaker Appearance Ratio | 0.24 | 0.00 | 0.62 | 0.49 |
| Frequency of faces per Shot | 0.05 | 0.25 | 0.22 | 0.34 |
| Frequency of faces per Minute | -0.08 | 0.05 | 1.55 | 2.41 |
| Frequency of objects | 0.29 | 0.00 | 13.39 | 12.68 |
| Frequency of Objects per Minute | -0.06 | 0.11 | 4.90 | 5.39 |
| Frequency of Objects per Shot | 0.17 | 0.00 | 1.89 | 1.40 |
| Frequency of Shots per Minute | 0.03 | 0.46 | 6.68 | 6.52 |
| Frequency of speakers per Minute | -0.21 | 0.00 | 1.89 | 2.73 |
| Frequency of speakers per Shot | -0.38 | 0.00 | 0.65 | 0.89 |
| Clutter FC (sum) | 0.23 | 0.00 | 145.29 | 246.72 |
| Clutter FC (mean) | 0.06 | 0.13 | 3.17 | 1.06 |
| Clutter FC (SD) | 0.34 | 0.00 | 0.86 | 0.58 |
| Clutter SE (Sum) | 0.25 | 0.00 | 115.51 | 182.47 |
| Clutter SE (mean) | 0.24 | 0.00 | 2.50 | 0.57 |
| Clutter SE (SD) | 0.28 | 0.00 | 0.37 | 0.25 |
| Colorfulness 5 sec (mean) | 0.06 | 0.13 | 29.22 | 13.44 |
| Colorfulness 5 sec (SD) | 0.15 | 0.00 | 11.07 | 8.30 |
| Colorfulness 5 sec Adjacent (mean) | -0.04 | 0.37 | 0.25 | 1.99 |
| Colorfulness 5 sec Adjacent (SD) | 0.16 | 0.00 | 15.79 | 11.87 |
| SSI 5 sec (mean) | -0.28 | 0.00 | 0.58 | 0.21 |
| SSI 5 sec (SD) | 0.27 | 0.00 | 0.16 | 0.08 |
| Colorfulness 1 sec (mean) | 0.09 | 0.03 | 28.13 | 13.60 |
| Colorfulness 1 sec (SD) | 0.08 | 0.04 | 10.51 | 8.06 |
| Colorfulness 1 sec Adjacent (mean) | 0.01 | 0.71 | -0.01 | 0.45 |
| Colorfulness 1 sec Adjacent (SD) | 0.00 | 0.99 | 6.90 | 5.99 |
| SSI 1 sec (mean) | 0.08 | 0.03 | 0.75 | 0.18 |
| SSI 1 sec (SD) | 0.00 | 0.92 | 0.18 | 0.10 |
| Saliency FG (sum) | 0.25 | 0.00 | 110265.73 | 169464.50 |
| Saliency FG (mean) | -0.08 | 0.06 | 2636.18 | 358.81 |
| Saliency FG (SD) | 0.36 | 0.00 | 286.15 | 219.11 |

**Table A.4 continued from previous page**

| Feature | *r* | Sig. | *M* | *SD* |
|---|---|---|---|---|
| Saliency SR (sum) | 0.25 | 0.00 | 507776.50 | 781392.13 |
| SaliencySR_mean | 0.11 | 0.01 | 11540.24 | 2159.75 |
| Saliency SR (SD) | 0.30 | 0.00 | 1917.24 | 1193.16 |
| Saliency IK (sum) | 0.25 | 0.00 | 583373.88 | 879020.56 |
| Saliency IK (mean) | 0.12 | 0.00 | 13559.19 | 1484.07 |
| Saliency IK (SD) | 0.28 | 0.00 | 1327.19 | 878.54 |
| KeyFrame BE1 (sum) | 0.22 | 0.00 | 914.00 | 1649.83 |
| KeyFrame KB (mean) | 0.03 | 0.45 | 19.15 | 7.12 |
| KeyFrame KB (SD) | 0.28 | 0.00 | 5.89 | 4.09 |
| KeyFrame KB (range) | 0.34 | 0.00 | 23.64 | 18.07 |
| Frequncy of moving objects | 0.33 | 0.00 | 413.32 | 536.00 |
| Motion duration(ms) | 0.05 | 0.20 | 259.37 | 765.33 |
| Moving Objects per minute | 0.23 | 0.00 | 108.08 | 101.47 |
| Motion Duration Difference (mean) | 0.34 | 0.00 | 299.44 | 270.79 |
| Motion Duration Difference (variance) | 0.10 | 0.01 | 603603.03 | 1791869.17 |
| Motion Duration Difference (SD) | 0.29 | 0.00 | 538.64 | 560.33 |
| Canny KeyFrame KB (sum) | 0.22 | 0.00 | 1252.09 | 2227.82 |
| Canny KeyFrame KB (mean) | 0.08 | 0.03 | 25.28 | 9.83 |
| Canny KeyFrame KB (SD) | 0.10 | 0.01 | 10.38 | 5.40 |
| Canny KeyFrame KB (range) | 0.29 | 0.00 | 34.61 | 25.26 |
| Canny video size (MB) | 0.30 | 0.00 | 130.07 | 123.85 |
| Video H264 size(MB) | 0.37 | 0.00 | 12.63 | 15.31 |
| Video H265_size(MB) | 0.30 | 0.00 | 2.94 | 3.89 |
| H265_size V30(MB) | 0.22 | 0.00 | 0.46 | 0.33 |
| Colorfulness V30 (mean) | 0.09 | 0.02 | 28.35 | 15.06 |
| Colorfulness V30 (SD) | 0.07 | 0.08 | 6.41 | 7.66 |
| Colorfulness Adjacent V30 (mean) | -0.02 | 0.59 | -0.06 | 0.89 |
| Colorfulness Adjacent V30 (SD) | 0.07 | 0.07 | 5.13 | 6.33 |
| SSI V30 (mean) | -0.20 | 0.00 | 0.75 | 0.20 |
| SSI V30 (SD) | 0.18 | 0.00 | 0.14 | 0.11 |
| Canny H264 V30( MB) | -0.06 | 0.14 | 7.88 | 4.67 |
| V30 H264 size(MB) | 0.26 | 0.00 | 1.29 | 0.87 |
| Number of moving objects V30 | 0.16 | 0.00 | 42.84 | 38.25 |
| Motion duration(ms) V30 | 0.08 | 0.03 | 327.29 | 875.53 |
| Moving Objects per second (V30) | 0.14 | 0.00 | 1.44 | 1.31 |
| Motion Duration Difference V30 (mean) | 0.28 | 0.00 | 300.39 | 394.85 |
| Motion Duration Difference V30 (variance) | 0.11 | 0.01 | 587289.38 | 2057855.26 |
| Motion Duration Difference V30 (SD) | 0.22 | 0.00 | 439.19 | 628.50 |

# Bibliography


Abdulhussain, S. H., Ramli, A. R., Saripan, M. I., Mahmmod, B. M., Al-Haddad, S. A. R., & Jassim, W. A. (2018). Methods and challenges in shot boundary detection: A review. *Entropy*, *20*(4). doi: 10.3390/e20040214

Abeywickrama, P. (2013). Why not non-native varieties of english as listening comprehension test input? *RELC Journal: A Journal of Language Teaching and Research*, *44*(1), 59–74. doi: 10.1177/0033688212473270

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 520–529. doi: 10.1037/a0013552

Adhikari, R., & Agrawal, R. K. (2013). An introductory study on time series modeling and forecasting. *arXiv preprint arXiv:1302.6613*.

Akramullah, S. (2014). *Digital video concepts, methods, and metrics: quality, compression, performance, and power trade-off analysis*. Springer Nature.

Albrecht, J. E., & Myers, J. L. (1995). Role of context in accessing distant information during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(6), 1459–1468. doi: 10.1037/0278-7393.21.6.1459

Albrecht, J. E., & O'Brien, E. J. (1993). Updating a mental model: Maintaining both local and global coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1061–1070. doi: 10.1037/0278-7393.19.5.1061

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*(4), 659–663.

Aldukhayel, D. (2019). Vlogs in l2 listening: Efl learners' and teachers' perceptions. *Computer Assisted Language Learning*, 1–20.

Alghamdi, E. A. (2018). *Video complexity and language learners: Making the match.* University of Sydney, Sydney, Australia: Presented at The Third Saudi Scientific Symbolism.

Alghamdi, E. A. (2019). *Video difficulty assessment.* School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia: Presented at the Human-Computer Interaction Seminars.

Alghamdi, E. A., Gruba, P., Marsi, A., & Velloso, E. (In Review). The use of lexical complexity for assessing difficulty in instructional videos. *Language Learning and Technology*.

Alghamdi, E. A., Gruba, P., & Velloso, E. (In Review). The relative contribution of verbal complexity to second language video lectures difficulty assessment. *Modern Language Journal*.

Alghamdi, E. A., Otaif, F., & Gruba, P. a. (2018). Designing a video playing interface for second language learners. In *Ascilite 2018* (pp. 298–302). doi: 10.1111/j.1365-2729.2011.00476.x

Alghamdi, E. A., Velloso, E., & Gruba, P. (2021). AUVANA: An automated video analysis tool for visual complexity.
doi: https://doi.org/10.31219/osf.io/kj9hx

Allen, G., & Hawkins, S. (1980). Phonological rhythm: Definition and development. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology: Volume 1. production* (pp. 227 – 256). San Diego, CA: Academic Press.

Alpaydin, E. (2018). Classifying multimodal data. In *The handbook of multimodal-multisensor interfaces: Foundations, user modeling, and common modality combinations - volume 2* (Vol. 2, pp. 49–69). doi: 10.1145/3107990.3107994

Ambard, P. D., & Ambard, L. K. (2012). Effects of narrative script advance organizer strategies used to introduce video in the foreign language classroom. *Foreign Language Annals*, *45*(2), 203–228. doi: 10.1111/j.1944-9720.2012.01189.x

Amendum, S. J., Conradi, K., & Hiebert, E. (2017). Does text complexity matter in the elementary grades? A research synthesis of text difficulty and elementary students' reading fluency and comprehension. *Educational Psychology Review*, 1–31. doi: 10.1007/s10648-017-9398-2

Anderson, J. (1996). *The reality of illusion: An ecological approach to cognitive film theory.* SIU Press.

Anderson, R. C., & Pichert, J. W. (1978). Recall of previously unrecallable information following a shift in perspective. *Journal of verbal learning and verbal behavior*, *17*(1), 1–12.

Andreessen, L. M., Gerjets, P., Meurers, D., & Zander, T. (2020). Toward neuroadaptive support technologies for improving digital reading: a passive bci-based assessment

of mental workload imposed by text difficulty and presentation speed during reading. *User Modeling and User-Adapted Interaction*, 1–30.

Arndt, H. L., & Woore, R. (2018). Vocabulary learning from watching youtube videos and reading blog posts. *Language Learning & Technology*, *22*(3), 124–142.

Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research*, *32*(1), 25–36. doi: 10.1023/a:1021980931292

Arnold, J. E., Kam, C., & Tanenhaus, M. (2007). 'If you say thee uh you are describing something hard': the online attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 914–30.

Arnold, T., Ballier, N., Gaillat, T., & Lissòn, P. (2018). Predicting CEFRL levels in learner English on the basis of metrics and full texts.

Aryadoust, V. (2019). An integrated cognitive theory of comprehension. *International Journal of Listening*(November), 1–30. doi: 10.1080/10904018.2017.1397519

Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, *16*(6), 345–379. doi: 10.1007/s00530-010-0182-0

Ausubel, D. P. (1960). The use of advance organizers in the learning and retention of meaningful verbal material. *Journal of educational psychology*, *51*(5), 267.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (release 2). *Distributed by the linguistic data consortium, University of Pennsylvania.*

Bailey, H. R., Kurby, C. A., Sargent, J. Q., & Zacks, J. M. (2017). Attentional focus affects how events are segmented and updated in narrative reading. *Memory and Cognition*, *45*(6), 940–955. doi: 10.3758/s13421-017-0707-2

Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language and Communication*, *21*(3), 285–301. doi: 10.1016/s0271-5309(01)00005-2

Bailin, A., & Grafstein, A. (2016). *Readability: Text and context* (Vol. 28) (No. 3). London: Palgrave Macmillan UK. doi: 10.1057/9781137388773

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? the role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human perception and performance*, *10*(3), 340.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The english lexicon project. *Behavior Research Methods*, *39*(3), 445–459. doi: 10.3758/bf03193014

Baltrusaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443. doi: 10.1109/tpami.2018.2798607

Balyan, R., McCarthy, K. S., & McNamara, D. S. (2020). Applying natural language processing and hierarchical machine learning approaches to text difficulty classification. *International Journal of Artificial Intelligence in Education*, 1–34.

Barber, H. A., Otten, L. J., Kousta, S.-T., & Vigliocco, G. (2013). Concreteness in word processing: Erp and behavioral effects in a lexical decision task. *Brain and language*, *125*(1), 47–53.

Batty, A. O. (2014). A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing*, *32*(1), 3–20. doi: 10.1177/0265532214531254

Batty, A. O. (2020). An eye-tracking study of attention to visual cues in l2 listening tests. *Language Testing*, 0265532220951504.

Beach, C. M. (1991). The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. *Journal of memory and language*, *30*(6), 644–663.

Becker, S. R., & Sturm, J. L. (2017). Effects of Audiovisual Media on L2 Listening Comprehension: A Preliminary Study in French. *CALICO Journal*, *34*(2), 147–177. doi: http://dx.doi.org/10.1558/cj.26754

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework* (Tech. Rep.). Princeton, NJ.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Benjamin, R. G. (2012). *Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty* (Vol. 24) (No. 1). doi: 10.1007/s10648-011-9181-8

Berger, K. W. (1977). *The most common 100,000 words used in conversations.* Herald Publishing House.

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics and Data Analysis*, *120*, 70–83. doi: 10.1016/j.csda.2017.11.003

Best, R., Floyd, R. G., & Mcnamara, D. S. (2004). Understanding the fourth-grade slump: Comprehension difficulties as a function of reader aptitudes and text genre. In *In 85th annual meeting of the american educational research association.*

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

Bhardwaj, S., & Mittal, A. (2012). A Survey on various edge detector techniques. *Procedia Technology*, *4*, 220–226. doi: 10.1016/j.protcy.2012.05.033

Biard, N., Cojean, S., & Jamet, E. (2018). Effects of segmentation and pacing on procedural learning by video. *Computers in Human Behavior*. doi: 10.1016/j.chb.2017.12.002

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

Biber, D. (1988). *Variation across speech and writing* (Vol. 3). doi: 10.1017/cbo9780511621024

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

Bjork, E. L., & Bjork, R. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York: Worth Publishers.

Bjork, R. (1994). Memory and metamemory considerations in the training ofhuman beings. In *Metacognition: Knowing about knowing* (p. 185–205). MIT Press.

Bläsing, B. E. (2014). Segmentation of dance movement: Effects of expertise, visual familiarity, motor experience and music. *Frontiers in Psychology*, *5*(Oct), 1–11. doi: 10.3389/fpsyg.2014.01500

Blau, E. K. (1990). The effect of syntax, speed, and pauses on listening comprehension. *TESOL Quarterly*, *24*(4), 746. doi: 10.2307/3587129

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., & Edu, J. B. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2010). What makes listening difficult? *University Of Maryland*(April), 92.

Boersma, P., & Weenink, D. (2019). *Praat: doing phonetics by computer.* http://www.praat.org/.

Boggia, J., & Ristic, J. (2015). Social event segmentation. *Quarterly Journal of Experimental Psychology*, *68*(4), 731–744. doi: 10.1080/17470218.2014.964738

Bordwell, D., & Thompson, K. (2001). *Film Art: An Introduction.* New York: McGraw-Hill. doi: film;culturalstudies;semiotik

Borji, A. (2015). What is a salient object? A dataset and a baseline model for salient object

detection. *IEEE Transactions on Image Processing*, *24*(2), 742–756. doi: 10.1109/tip.2014.2383320

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207. doi: 10.1109/tpami.2012.89

Bormuth, J. R. (1968). Cloze test readability: Criterion reference scores. *Journal of educational measurement*, *5*(3), 189–196.

Bozorgian, H., & Alamdari, E. F. (2018). Multimedia listening comprehension: Metacognitive instruction or metacognitive instruction through dialogic interaction. *ReCALL: the Journal of EUROCALL*, *30*(1), 131.

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*(4 Pt 1), 2074–85. doi: 10.1121/1.427952

Bramão, I., Reis, A., Petersson, K. M., & Faísca, L. (2011). The role of color information on object recognition: A review and meta-analysis. *Acta Psychologica*, *138*(1), 244–253. doi: 10.1016/j.actpsy.2011.06.010

Braun, J., Amirshahi, S. A., Denzler, J., & Redies, C. (2013). Statistical image properties of print advertisements, visual artworks and images of architecture. *Frontiers in Psychology*, *4*, 808.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, *19*(4), 369–394. doi: 10.1191/0265532202lt236oa

Briscoe, T., Carroll, J., & Watson, R. (2006). The Second Release of the RASP System. *Proceedings of the COLING/ACL on Interactive presentation sessions.*(July), 77–80. doi: 10.3115/1225403.1225423

Britton, B. K., & Gülgöz, S. (1991). Using kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of educational Psychology*, *83*(3), 329.

Broadbent, D. (1958). *Perception and communication.* Pergamon Press.

Broth, M., Laurier, E., & Mondada, L. (2014). Introducing video at work. In M. Broth, E. Laurier, & L. Mondada (Eds.), *Studies of video practices: Video at work.* Routledge.

Brown, J. D. (1998). An EFL Readability Index. *JALT Journal*, *20*(2), 7–36.

Brown, J. D. (2012). *New ways in teaching connected speech.* Alexandria, VA: TESOL International Association.

Brunfaut, T., & Révész, A. (2015). The role of task and listener characteristics in second language listening. *TESOL Quarterly*, *49*(1), 141–168. doi: 10.1002/tesq.168

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. doi: 10.3758/s13428-013-0403-5

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*(2), 119–157. doi: 10.1177/026553229801500201

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In *Dimensions of l2 performance and proficiency: Complexity; accuracy and fluency in sla* (pp. 21–46). doi: 10.1075/lllt.32.02bul

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In *Handbook of automated essay evaluation: Current applications and new directions* (pp. 55–67).

Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *Pami-8*(6), 679–698. doi: 10.1109/tpami.1986.4767851

Cardaci, M., Di Gesù, V., Petrou, M., & Tabacchi, M. E. (2009). A fuzzy approach to the evaluation of image complexity. *Fuzzy Sets and Systems*, *160*(10), 1474–1484. doi: 10.1016/j.fss.2008.11.017

Cárdenas-Claros, M. S. (2020). Spontaneous links between help option use and input features that hinder second language listening comprehension. *System*, *93*, 102308.

Carrasco, M. (2011). Visual attention: the past 25 years. *Vision research*, *51*(13), 1484–525. doi: 10.1016/j.visres.2011.04.012

Carver, R. P. (1985). Measuring readability using drp units. *Journal of Reading Behavior*, *17*(4), 303–316.

Cassell, J., McNeill, D., & McCullough, K.-E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & cognition*, *7*(1), 1–34.

Cerqueira, V., Torgo, L., & Mozetic, I. (2019). Evaluating time series forecasting models: An empirical study on performance estimation methods. , 1–28.

Cervantes, R., & Gainer, G. (1992). The Effects of Syntactic Simplification and Repetition on Listening Comprehension. *TESOL Quarterly*, *26*(4), 767. doi: 10.2307/3586886

Cervetti, G. N., Bravo, M. A., Hiebert, E. H., Pearson, P. D., & Jaynes, C. A. (2009). Text genre and science content: Ease of reading, comprehension, and reader preference. *Reading Psychology*, *30*(6), 487–511. doi: 10.1080/02702710902733550

Cevasco, J., & van den Broek, P. (2016). The effect of filled pauses on the processing of

the surface form and the establishment of causal connections during the comprehension of spoken expository discourse. *Cognitive Processing*, *17*(2), 185–194. doi: 10.1007/s10339-016-0755-8

Chall, J. S. (1958). *Readability: An appraisal of research and application* (No. 34). Ohio State University.

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new dale-chall readability formula*. Brookline Books.

Chang, A. C.-S. (2018). Speech rate second language listening. *The TESOL Encyclopedia of English Language Teaching*, *1968*, 1–7. doi: 10.1002/9781118784235.eelt0580

Chatfield, C. (2000). *Time-series forecasting*. Chapman and Hall/CRC.

Chen, C.-M., Li, M.-C., & Lin, M.-F. (2020). The effects of video-annotated learning and reviewing system with vocabulary learning mechanism on english listening comprehension and technology acceptance. *Computer Assisted Language Learning*, 1–37.

Chen, D., & Manning, C. (2015). A fast and accurate dependency parser using neural networks. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*(i), 740–750. doi: 10.3115/v1/d14-1082

Chen, H. J. H. (2011). Developing and evaluating SynctoLearn, a fully automatic video and transcript synchronization tool for EFL learners. *Computer Assisted Language Learning*, *24*(2), 117–130. doi: 10.1080/09588221.2010.526947

Chen, L., Zechner, K., Yoon, S.-Y., Evanini, K., Wang, X., Loukina, A., … Gyawali, B. (2018). Automated Scoring of Nonnative Speech Using the SpeechRater SM v. 5.0 Engine. *ETS Research Report Series*(April). doi: 10.1002/ets2.12198

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1724–1734. doi: 10.3115/v1/d14-1179

Chotlos, J. (1944). IV. A statistical and comparative analysis of individual written language samples. *Psychological Monographs*, *56*(2), 75–111.

Chung, J. (1999). The effects of using video texts supported with advanced organizers and captions on Chinese college students' listening comprehension: An empirical study. *Foreign Language Annals*, *32*(3), 295–308.

Chung, J. (2002). The effects of using two advance organizers with video texts for the teaching of listening in english. *Foreign Language Annals*, *35*(2), 231–241.

Chung, J., & Huang, S. (1998). The effects of three aural advance organizers for video

viewing in a foreign language classroom. *System*, *26*(4), 553–565. doi: 10.1016/s0346-251x(98)00037-2

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Cintrón-Valentín, M., García-Amaya, L., & Ellis, N. C. (2019). Captioning and grammar learning in the l2 spanish classroom. *The Language Learning Journal*, *47*(4), 439–459.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge university press. doi: 10.1016/s0378-2166(97)83330-9

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*(3), 149–210. doi: 10.1007/bf01320076

Clarke, A. D., Coco, M. I., & Keller, F. (2013). The impact of attentional, linguistic, and visual features during object naming. *Frontiers in Psychology*, *4*(Dec), 1–12. doi: 10.3389/fpsyg.2013.00927

Coco, M. I., & Keller, F. (2012). Scan Patterns Predict Sentence Production in the Cross-Modal Processing of Visual Scenes. *Cognitive Science*, *36*(7), 1204–1223. doi: 10.1111/j.1551-6709.2012.01246.x

Coco, M. I., Keller, F., & Malcolm, G. L. (2016). Anticipation in Real-World Scenes: The Role of Visual Context and Visual Memory. *Cognitive Science*, *40*(8), 1995–2024. doi: 10.1111/cogs.12313

Coco, M. I., Malcolm, G. L., & Keller, F. (2014). The interplay of bottom-up and top-down mechanisms in visual guidance during object naming. *Quarterly Journal of Experimental Psychology*, *67*(6), 1096–1120. doi: 10.1080/17470218.2013.844843

Cohen, M. A., & Rubenstein, J. (2020). How much color do we see in the blink of an eye? *Cognition*, *200*, 104268.

Cohen Priva, U. (2017). Not so fast: Fast speech correlates with lower lexical and structural information. *Cognition*, *160*, 27–34. doi: 10.1016/j.cognition.2016.12.002

Cohn, N. (2013). Visual Narrative Structure. *Cognitive Science*, *37*(3), 413–452. doi: 10.1111/cogs.12016

Cohn, N. (2016a). From visual narrative grammar to filmic narrative grammar the narrative structure of static and moving images. *Film Text Analysis: New Perspectives on the Analysis of Filmic Meaning*, 94–117. doi: 10.4324/9781315692746

Cohn, N. (2016b). A multimodal parallel architecture: A cognitive framework for multimodal interactions. *Cognition*, *146*, 304–323.

Cohn, N. (2020). Your brain on comics: A cognitive model of visual narrative comprehension. *Topics in cognitive science*, *12*(1), 352–386.

Collard, P., Corley, M., MacGregor, L. J., & Donaldson, D. I. (2008). Attention Orienting Effects of Hesitations in Speech: Evidence From ERPs. *Journal of Experimental Psychology: Learning Memory and Cognition*, *34*(3), 696–702. doi: 10.1037/0278-7393.34.3.696

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of past, present, and future research. *Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics*(165:2), 97–135. doi: 10.1075/itl.165.2.01col

Collins-Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, *56*(13), 1448–1462. doi: 10.1002/asi.20243

Coltheart, M. (1981). The MRC Psycholinguistic Database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505. doi: 10.1080/14640748108400805

Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study. *System*, *29*(1), 1–14. doi: 10.1016/s0346-251x(00)00057-9

Connine, C. M., Mullennix, J., Shernoff, E., & Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(6), 1084.

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

Corchs, S., Ciocca, G., Bricolo, E., & Gasparini, F. (2016). Predicting complexity perception of real world images. *PLoS ONE*, *11*(6). doi: 10.1371/journal.pone.0157986

Corley, M., & Hartsuiker, R. J. (2003). Hesitation in speech can. . . um. . . help a listener understand. In *Proceedings of the annual meeting of the cognitive science behaviour* (pp. 276–281). doi: https://doi.org/ISBN978-0-9768318-8-4

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

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

Crook, C., & Schofield, L. (2017). The video lecture. *Internet and Higher Education*, *34*(January), 56–64. doi: 10.1016/j.iheduc.2017.05.003

Cross, J. (2009). Effects of listening strategy instruction on news videotext comprehension. *Language Teaching Research*, *13*(2), 151–176. doi: 10.1177/1362168809103446

Cross, J. (2011). Comprehending news videotexts: The influence of the visual content. *Language Learning & Technology*, *15*(2), 44–68.

Cross, J. (2018). Video in Listening. In *The tesol encyclopedia of english language teaching* (pp. 1–6). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi: 10.1002/9781118784235.eelt0606

Crossley, S. A., Allen, D. B., & McNamara, D. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, *23*(1), 84–101.

Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System*, *41*(4), 965–981. doi: 10.1016/j.system.2013.08.002

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, *42*(3), 475 – 493. doi: 10.1002/j.1545-7249.2008.tb001

Crossley, S. A., Kyle, K., & Dascalu, M. (2018). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, 14–27. doi: 10.3758/s13428-018-1142-4

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, *51*(1), 14–27. doi: 10.3758/s13428-018-1142-4

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*(4), 1227–1237. doi: 10.3758/s13428-015-0651-7

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115–135. doi: 10.1111/j.1467-9817.2010.01449.x

Crossley, S. A., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, *59*(2), 307–334. doi: 10.1111/j.1467-9922.2009.00508.x

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580. doi: 10.1177/0265532210378031

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, *42*(3-4), 541–561. doi: 10.1111/1467-9817.12283

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting

text comprehension, processing, and familiarity in adult Readers: New approaches to readability formulas. *Discourse Processes*, *54*(5-6), 340–359. doi: 10.1080/0163853x .2017.1296264

Crystal, D. (1994). *An encyclopedic dictionary of language and languages*. Penguin.

Cubilo, J., & Winke, P. (2013). Redefining the L2 Listening Construct Within an Integrated Writing Task: Considering the Impacts of Visual-Cue Interpretation and Note-Taking. *Language Assessment Quarterly*, *10*(4), 371–397. doi: 10.1080/ 15434303.2013.824972

Cunningham, J. W., & Mesmer, H. A. (2014). Quantitative measurement of text difficulty: What's the Use? *Elementary School Journal*, *115*(2), 255–269. doi: 10.1086/678292

Cunnings, I. (2017). Parsing and working memory in bilingual sentence processing. *Bilingualism: Language and Cognition*, *20*(4), 659–678.

Curto, P., Mamede, N., & Baptista, J. (2015). Automatic Text Difficulty Classifier:Assisting the Selection Of Adequate Reading Materials For European Portuguese Teaching. In *7th international conference on computer supported education 4* (pp. 36–44).

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and speech*, *40*(2), 141–201.

Cutler, A., & Otake, T. (1994). *Mora or Phoneme? Further Evidence for Language-Specific Listening* (Vol. 33) (No. 6). doi: 10.1006/jmla.1994.1039

Cutting, J. E., DeLong, J. E., & Brunick, K. L. (2011). Visual activity in hollywood film: 1935 to 2005 and beyond. *Psychology of Aesthetics, Creativity, and the Arts*, *5*(2), 115.

Dahl, T. I., & Ludvigsen, S. (2014). How I see what you're saying: The role of gestures in native and foreign language listening comprehension. *The Modern Language Journal*, *98*(3), 813–833.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, 37–54.

Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English*, *26*(1), 19–26.

Dale, E., & Tyler, R. W. (1934). A study of the factors influencing the difficulty of reading materials for adults of limited reading ability. *The Library Quarterly*, *4*(3), 384–412.

Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and Assessing Vocabulary Knowledge* (H. Daller, J. Milton, & J. Treffers-Daller, Eds.). Cambridge: Cambridge University Press. doi: 10.1017/cbo9780511667268

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The Academic Spoken Word List. *Language Learning*, *67*(4), 959–997. doi: 10.1111/lang.12253

Dargue, N., & Sweller, N. (2018). Not all gestures are created equal: The effects of typical and atypical iconic gestures on narrative comprehension. *Journal of Nonverbal*

*Behavior*, *42*(3), 327–345.

Dargue, N., & Sweller, N. (2020a). Learning stories through gesture: Gesture's effects on child and adult narrative comprehension. *Educational Psychology Review*, *32*(1), 249–276.

Dargue, N., & Sweller, N. (2020b). Two hands and a tale: When gestures benefit adult narrative comprehension. *Learning and Instruction*, *68*, 101331.

Dargue, N., Sweller, N., & Jones, M. P. (2019). When our hands help us understand: A meta-analysis into the effects of gesture on comprehension. *Psychological Bulletin*, *145*(8), 765.

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 447–464. doi: 10.1093/llc/fqq018

Davison, A., & Kantor, R. (1982). On the Failure of Readability Formulas to Define Readable Texts: A Case Study from Adaptations. *Reading Research Quarterly*, *17*(2), 187–209.

De Clercq, O., Hoste, V., Desmet, B., Van Oosten, P., De Cock, M., & Macken, L. (2014). Using the crowd for readability prediction. *Natural Language Engineering*, *20*(03), 293–325. doi: 10.1017/s1351324912000344

De Koning, B. B., Tabbers, H. K., Rikers, R. M., & Paas, F. (2009). Towards a Framework for Attention Cueing in Instructional Animations: Guidelines for Research and Design. *Educational Psychology Review*, *21*(2), 113–140. doi: 10.1007/s10648-009-9098-7

De Tommaso, M., & Turatto, M. (2019). Learning to ignore salient distractors: Attentional set and habituation. *Visual Cognition*, *6285*. doi: 10.1080/13506285.2019.1583298

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390. doi: 10.3758/brm.41.2.385

DeKeyser, R. (2016). Of moving targets and chameleons. *Studies in Second Language Acquisition*, *38*(2), 353–363. doi: 10.1017/s0272263116000024

DeKeyser, R. M. (2013). Age effects in second language learning: Stepping stones toward better understanding. *Language Learning*, *63*, 52–67.

Dell'Orletta, F., Venturi, G., & Montemagni, S. (2012). Genre–oriented Readability Assessment: A Case Study. *COLING (Workshop)*(December), 91–98.

DeSchepper, B., & Treisman, A. (1996). Visual memory for novel shapes: implicit coding without attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 27.

Deutsch, T., Jasbi, M., & Shieber, S. (2020). Linguistic features for readability assessment. In *Proceedings of the fifteenth workshop on innovative use of nlp for building educational applications* (pp. 1–17). Seattle, WA, USA â†' Online: Association for Computational Linguistics.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78. doi: 10.1145/2347736.2347755

Donderi, D. C. (2006). Visual complexity: a review. *Psychological bulletin*, *132*(1), 73.

Donderi, D. C., & McFadden, S. (2005). Compressed file length predicts search time and errors on visual displays. *Displays*, *26*(2), 71–78.

Dörnyei, Z. (2014). *The psychology of the language learner: Individual differences in second language acquisition*. Routledge.

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of vision*, *10*(10), 28–28.

Douglas, H. E. (2009). Reintroducing prediction to explanation. *Philosophy of Science*, *76*(4), 444–463.

Douglas Fir Group. (2016). A Transdisciplinary Framework for SLA in a Multilingual World. *Modern Language Journal*, *100*, 19–47. doi: 10.1111/modl.12301

Drijvers, L., Vaitonytė, J., & Özyürek, A. (2019). Degree of language experience modulates visual attention to visible speech and iconic gestures during clear and degraded speech comprehension. *Cognitive science*, *43*(10), e12789.

DuBay, W. H. (2004). The principles of readability: A brief introduction to readability research. *Costa Mesa: Impact Information*, *76*.

DuBay, W. H. (2007). *Smart language: Readers, readability, and the grading of text.* ERIC.

Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics*, *25*(2), 220–242+287. doi: 10.1093/applin/25.2.220

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 49-98.

Ellis, N. C. (2002). Frequency effects in language processing. A Review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*(24), 143–188. doi: 10.1017.s0272263102002024

Ellis, N. C., Simpson-vlach, R., & Maynard, C. (2008). Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL. *TESOL Quarterly*, *42*(3), 375–396. doi: 10.1002/j.1545-7249.2008.tb00137.x

Ernestus, M., Kouwenhoven, H., & van Mulken, M. (2017). The direct and indirect effects of the phonotactic constraints in the listener's native language on the comprehension of reduced and unreduced word pronunciation variants in a foreign language.

*Journal of Phonetics*, *62*, 50–64. doi: 10.1016/j.wocn.2017.02.003

Evans, K. K., Horowitz, T. S., Howe, P., Pedersini, R., Reijnen, E., Pinto, Y., … Wolfe, J. M. (2011). Visual attention. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 503–514. doi: 10.1002/wcs.127

Feng, L., Elhadad, N., & Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. *EACL 2009 - 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*(April), 229–237. doi: 10.3115/1609067.1609092

Feng, Y., & Webb, S. (2019). Learning vocabulary through reading, listening, and viewing: which mode of input is most effective? *Studies in Second Language Acquisition*, 1–25.

Feng, Y., & Webb, S. (2020). Learning vocabulary through reading, listening, and viewing: Which mode of input is most effective? *Studies in Second Language Acquisition*, *42*(3), 499–523.

Ferreira, F., Foucart, A., & Engelhardt, P. E. (2013). Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of Memory and Language*, *69*(3), 165–182. doi: 10.1016/j.jml.2013.06.001

Ferstl, E. C. (2018). Text comprehension. In *The oxford handbook of psycholinguistics* (pp. 197–216).

FFmpeg Developers. (2016). *ffmpeg*.

Field, J. (2003). Promoting perception: Lexical segmentation in l2 listening. *ELT journal*, *57*(4), 325–334.

Filighera, A., Steuer, T., & Rensing, C. (2019). Automatic Text Difficulty Estimation Using Embeddings and Neural Networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11722 Lncs*, 335–348. doi: 10.1007/978-3-030-29736-7_25

Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing*. Oxford University Press.

Fiorella, L., & Mayer, R. E. (2018). What works and doesn't work with instructional video. *Computers in Human Behavior*, *89*(July), 465–470. doi: 10.1016/j.chb.2018.07.015

Fiorella, L., Stull, A. T., Kuhlmann, S., & Mayer, R. E. (2019). Fostering generative learning from video lessons: Benefits of instructor-generated drawings and learner-generated explanations. *Journal of Educational Psychology*. doi: 10.1037/edu0000408

Fiorella, L., Van Gog, T., Hoogerheide, V., & Mayer, R. E. (2017). It's all a matter of perspective: Viewing first-person video modeling examples promotes learning of

an assembly task. *Journal of Educational Psychology*, *109*(5), 653–665. doi: 10.1037/edu0000161

Fisher, D., Frey, N., & Lapp, D. (2016). *Text complexity: Stretching readers with texts and tasks*. Corwin Press.

Flesch, R. (1943). Marks of readable style: Study in adult education. *Teachers College Contributions to Education*.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221–233.

Flesch, R. (1949). *The art of readable writing*. Harper New York.

Floccia, C., Butler, J., Goslin, J., & Ellis, L. (2009). Regional and foreign accent processing in English: Can listeners adapt? *Journal of Psycholinguistic Research*, *38*(4), 379–412. doi: 10.1007/s10936-008-9097-8

Flores, S., Bailey, H. R., Eisenberg, M. L., & Zacks, J. M. (2017). Event segmentation improves event memory up to one monthsss later. *Journal of Experimental Psychology: Learning Memory and Cognition*, *43*(8), 1183–1202. doi: 10.1037/xlm0000367

Flowerdew, J., & Miller, L. (1992). Student Perceptions, Problems and Strategies in Second Language Lecture Comprehension. *RELC Journal*, *23*(2), 60–80.

Flowerdew, J., & Miller, L. (2005). *Second language listening: Theory and practice*. Cambridge University Press.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, *25*(2-3), 285–307.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375. doi: 10.1093/applin/21.3.354

Frintrop, S., Rome, E., & Christensen, H. I. (2010). Computational visual attention systems and their cognitive foundations. *ACM Transactions on Applied Perception*, *7*(1), 1–39. doi: 10.1145/1658349.1658355

Fulcher, G. (1997). Text difficulty and accessibility: Reading formulae and expert judgement. *System*, *25*(4), 497–513.

García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining* (Vol. 72). Cham: Springer International Publishing. doi: 10.1007/978-3-319-10247-4

Gardner, H. (1987). *The mind's new science: A history of the cognitive revolution*. New York: Basic books.

Gartus, A., & Leder, H. (2017). Predicting perceived visual complexity of abstract patterns using computational measures: The influence of mirror symmetry on complexity perception. *PLoS ONE*, *12*(11), 1–30. doi: 10.1371/journal.pone.0185276

Gaspelin, N., & Luck, S. J. (2018). The Role of Inhibition in Avoiding Distraction by Salient Stimuli. *Trends in Cognitive Sciences*, *22*(1), 79–92. doi: 10.1016/j.tics.2017.11.001

Gaspelin, N., & Luck, S. J. (2019). Inhibition as a potential resolution to the attentional capture debate. *Current Opinion in Psychology*, *29*, 12–18. doi: 10.1016/j.copsyc.2018.10.013

Gass, S., Winke, P., Isbell, D. R., & Ahn, J. (2019). How captions help people learn languages: A working-memory, eye-tracking study.

Gernsbacher, M. A. (1983). *Memory for the orientation ofpictures in nonverbal stories: Parallels and insights into language processing* (Unpublished doctoral dissertation). Unpublished doctoral dissertation, University of Texas at Austin.

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.

Gernsbacher, M. A. (1997). Two decades of structure building. *Discourse Processes*, *23*(3), 265–304. doi: 10.1080/01638539709544994

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76. doi: 10.1016/s0010-0277(98)00034-1

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain* (pp. 95–126.).

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing*, *19*(2), 133–167. doi: 10.1191/0265532202lt225oa

Givón, T. (2001). *Syntax: An introduction*. John Benjamins Publishing.

Glanzer, M., Fischer, B., & Dorfman, D. (1984). Short-term storage in reading. *Journal of Verbal Learning and Verbal Behavior*, *23*(4), 467–486. doi: 10.1016/s0022-5371(84)90300-1

Gold, D. A., Zacks, J. M., & Flores, S. (2017). Effects of cues to event segmentation on subsequent memory. *Cognitive Research: Principles and Implications*, *2*(1), 1. doi: 10.1186/s41235-016-0043-2

Goldman, S. R., & Rakestraw Jr, J. A. (2000). Structural aspects of constructing meaning from text. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook ofreadingresearch (vol. 3)* (p. 311–336). Lawrence Erlbaum Associates Publishers.

Graesser, A. C., Hauft-Smith, K., Cohen, A. D., & Pyles, L. D. (1980). Advanced outlines, familiarity, and text genre on retention of prose. *Journal of Experimental Education*, *48*(4), 281–290. doi: 10.1080/00220973.1980.11011745

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do Readers Need to Learn in Order to Process Coherence Relations in Narrative and Expository Text? *Rethinking reading comprehension*, *3230*(901), 82–98. doi: 10.1017/cbo9781107415324.004

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis

of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, *36*(2), 193–202. doi: 10.3758/bf03195564

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*(3), 371–395. doi: 10.1037/0033-295x .101.3.371

Graf, I. (2008). Respondent burden. *Encyclopedia of Survey Research Methods*, 739–740. doi: 10.4135/9781412963947

Graham, S. (2006). Listening comprehension: The learners' perspective. *System*, *34*(2), 165–182. doi: 10.1016/j.system.2005.11.001

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of memory and language*, *29*(2), 228–244.

Granger, S., Gilquin, G., & Meunier, F. (2015). *The cambridge handbook of learner corpus research.* Cambridge University Press.

Green, C. (2015). An analysis of the relationship between cohesion and clause combination in English discourse employing NLP and data mining approaches. *Digital Scholarship in the Humanities*, *30*(3), 326–343. doi: 10.1093/llc/fqu012

Greenfield, G. (1999). *Classic readability formulas in an EFL context: Are they valid for Japanese speakers?* (Unpublished doctoral dissertation). Temple University, Philadelphia, PA, United States.

Greenfield, Jerry. (2004). Readability Formula for EFL. *Japan Association for Language Teacing Journal*, *26*(1), 5–24.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*(4), 403–437. doi: 10.1075/ijcl.13.4.02gri

Gries, S. T. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics*, *18*(1), 137–166. doi: 10.1075/ijcl.18.1.09gri

Gries, S. T., & Ellis, N. C. (2015). Statistical Measures for Usage-Based Linguistics. *Language Learning*, *65*(S1), 228–255. doi: 10.1111/lang.12119

Griffiths, R. (1990). Speech rate and non-native speaker comprehension: a preliminary study in time-benefit analysis. *Language Learning*, *40*(3), 311–336.

Griffiths, R. (1992). Speech Rate and Listening Comprehension: Further Evidence of the Relationship. *TESOL Quarterly*, *26*(2), 385–390.

Gruba, P. (1993). A comparison study of audio and video in language testing. *JALT journal*, *15*(1), 85–88.

Gruba, P. (2004). Understanding digitized second language videotext. *Computer Assisted Language Learning*, *17*(1), 51–82. doi: 10.1076/call.17.1.51.29710

Gruba, P. (2006). Playing the videotext: A media literacy perspective on video-mediated l2 listening. *Language Learning & Technology*, *10*(2), 77–92.

Guichon, N., & McLornan, S. (2008). The effects of multimodality on L2 learners: Implications for CALL resource design. *System*, *36*(1), 85–93. doi: 10.1016/j.system.2007.11.005

Gullberg, M. (2010). Methodological reflections on gesture analysis in second language acquisition and bilingualism research. *Second language research*, *26*(1), 75–102.

Gunning, R. (1952). Technique of clear writing.

Guo, W., Wang, J., & Wang, S. (2019). Deep multimodal representation learning: A survey. *IEEE Access*, *7*, 63373–63394. doi: 10.1109/access.2019.2916887

Gurban, M., Thiran, J.-P., Drugman, T., & Dutoit, T. (2008). Dynamic modality weighting for multi-stream hmms inaudio-visual speech recognition. In *Proceedings of the 10th international conference on multimodal interfaces* (pp. 237–240).

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*(2), 201. doi: 10.2307/3588378

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies.Association for Computational Linguistics*, 1–8. doi: 10.3115/1073336.1073357

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*(2), 8–12. doi: 10.1109/mis.2009.36

Hall, K. C., Allen, B., Fry, M., Khia Johnson, R. L., Mackie, S., & McAuliffe, M. (2017). *Phonological CorpusTools, Version 1.3.*

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

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* London: Longman.

Halliday, M. A. K., & Hasan, R. (1985). *Language, context and text: Aspects of language in a social-semiotic perspective.* Victoria: Deakin university press.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's Introduction to Functional Grammar.* Routledge.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques.* Elsevier.

Hansch, A., Hillers, L., McConachie, K., Newman, C., Schildhauer, T., Schmidt, J. P., & Schmidt, P. (2015). Video and Online Learning: Critical Reflections and Findings from the Field. *Ssrn.* doi: 10.2139/ssrn.2577882

Hashimoto, B. J., & Egbert, J. (2019). More Than Frequency? Exploring Predictors of Word Difficulty for Second Language Learners. *Language Learning*, 1–34. doi:

10.1111/lang.12353

Hasler, D., & Suesstrunk, S. E. (2003). Measuring colorfulness in natural images. In B. E. Rogowitz & T. N. Pappas (Eds.), (p. 87). doi: 10.1117/12.477378

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. doi: 10.1007/b94608

Heaps, C., & Handel, S. (1999). Similarity and features of natural textures. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(2), 299–320. doi: 10.1037/0096-1523.25.2.299

Heaps, H. (1978). *Information retrieval: Computational and theoretical aspects*. New York: Academic Press.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In *Human language technologies 2007: The conference of the north american chapter of the association for computational linguistics; proceedings of the main conference* (pp. 460–467). doi: 10.1.1.70.1391

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*(2), 135–175. doi: 10.1086/286983

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual review of psychology*, *50*(1), 243–271.

Hendriks Vettehen, P., & Kleemans, M. (2015). How Camera Changes and Information Introduced Affect the Recognition of Public Service Announcements: A Test Outside the Lab. *Communication Research*, *46*(7), 908–925. doi: 10.1177/0093650215616458

Henrichsen, L. E. (1984). Sandhi[U+2010]Variation: a Filter of Input for Learners of Esl. *Language Learning*, *34*(3), 103–123. doi: 10.1111/j.1467-1770.1984.tb00343.x

Herron, C., Cole, S. P., Corrie, C., & Dubreil, S. (1999). The effectiveness of a video-based curriculum in teaching culture. *The Modern Language Journal*, *83*(4), 518–533.

Herron, C., Dubreil, B., Corrie, C., & Cole, S. P. (2002). A classroom investigation: can video improve intermediate-level french language students' ability to learn about a foreign culture? *The Modern Language Journal*, *86*(1), 36–53.

Herron, C., Dubreil, S., Cole, S. P., & Corrie, C. (2000). Using instructional video to teach culture to beginning foreign language students. *Calico Journal*, 395–429.

Herron, C., York, H., Corrie, C., & Cole, S. P. (2006). A comparison study of the effects of a story-based video instructional package versus a text-based instructional package in the intermediate-level foreign language classroom. *CALICO Journal*, *23*(2), 281–307. doi: 10.1558/cj.v23i2.281-307

Hershler, O., Golan, T., Bentin, S., & Hochstein, S. (2010). The wide window of face

detection Orit Hershler Shlomo Bentin. *Journal of Vision*, *10*(2010), 1–14. doi: 10 .1167/10.10.21.Introduction

Hiebert, E. H., & Pearson, P. D. (2014). Understanding Text Complexity: Introduction to the Special Issue. *The Elementary School Journal*, *115*(2), 153–160. doi: 10.1086/678446

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Holle, H., Obleser, J., Rueschemeyer, S.-A., & Gunter, T. C. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *Neuroimage*, *49*(1), 875–884.

Hoover, D. L. (2003). Another perspective on vocabulary richness. *Computers and the Humanities*, *37*(2), 151–178. doi: 10.1023/a:1022673822140

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. doi: 10.1016/0893-6080(89)90020-8

Hostetter, A. B. (2011). When do gestures communicate? a meta-analysis. *Psychological bulletin*, *137*(2), 297.

Hou, X., & Zhang, L. (2007). Saliency Detection: A Spectral Residual Approach. In *2007 ieee conference on computer vision and pattern recognition* (pp. 1–8). Ieee. doi: 10.1109/cvpr.2007.383267

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

Housen, A., & Simoens, H. (2016). Introduction: Cognitive perspectives on difficulty and complexity in L2 acquisition. *Studies in Second Language Acquisition*, *38*(2), 163–175. doi: 10.1017/s0272263116000176

Hsieh, Y. (2020). Effects of video captioning on EFL vocabulary learning and listening comprehension. *Computer Assisted Language Learning*, *0*(0), 1–23. doi: 10.1080/09588221.2019.1577898

Hunston, S. (2002). *Corpora in applied linguistics*. Ernst Klett Sprachen.

Hunt, K. (1965). *Grammatical Structures Written at Three Grade Levels*. Champaign, Ill: National Council of Teachers of English.

Hunt, K. (1970). Syntactic Maturity in Schoolchildren and Adults. *The Social History of the American Family: An Encyclopedia*, *35*(1). doi: 10.4135/9781452286143.n496

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.

Ildirar, S., & Schwan, S. (2015). First-time viewers' comprehension of films: Bridging shot transitions. *British Journal of Psychology*, *106*(1), 133–151. doi: 10.1111/bjop.12069

Itti, L., & Baldi, P. (2005). A principled approach to detecting surprising events in video. *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, *I*(June), 631–637. doi: 10.1109/cvpr.2005.40

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. doi: 10.1109/34.730558

Jaderberg, M., Vedaldi, A., & Zisserman, A. (2014). Deep features for text spotting. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8692 Lncs*(Part 4), 512–528. doi: 10.1007/978-3-319-10593-2_34

Jafari, K., & Hashim, F. (2012). The effects of using advance organizers on improving EFL learners' listening comprehension: A mixed method study. *System*, *40*(2), 270–281. doi: 10.1016/j.system.2012.04.009

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). New York, NY: Springer New York. doi: 10.1007/978-1-4614-7138-7

James, W. (2007). *The principles of psychology* (Vol. 1). Cosimo, Inc.

Jin, T., Lu, X., & Ni, J. (2020). Syntactic complexity in adapted teaching materials: Differences among grade levels and implications for benchmarking. *Modern Language Journal*. doi: 10.1111/modl.12622

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Lund Working Papers in Linguistics*, *53*, 61–79.

Johnson, A., & Proctor, R. W. (2004). *Attention: Theory and practice*. Sage.

Joiner, E. G. (1990). Choosing and using videotexts. *Foreign Language Annals*, *23*, 53–64.

KaewTraKulPong, P., & Bowden, R. (2002). An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection. *Video-Based Surveillance Systems*, 135–144. doi: 10.1007/978-1-4615-0913-4_11

Kang, O., Thomson, R., & Moran, M. (2018). The Effects of International Accents and Shared First Language on Listening Comprehension Tests. *TESOL Quarterly*, *53*(1), 56–81. doi: 10.1002/tesq.463

Karpov, N., Baranova, J., & Vitugin, F. (2014). Single-sentence readability prediction in Russian. In *Communications in computer and information science* (Vol. 436, pp. 91–100). doi: 10.1007/978-3-319-12580-0_9

Kate, R., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R., … Welty, C. (2010).

Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd international conference on computational linguistics (coling 2010)* (pp. 546–554).

Kavak, Y., Erdem, E., & Erdem, A. (2017). A comparative study for feature integration strategies in dynamic saliency estimation. *Signal Processing: Image Communication*, *51*(November 2016), 13–25. doi: 10.1016/j.image.2016.11.003

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260–267.

Kendeou, P., McMaster, K. L., Butterfuss, R., Kim, J., Bresina, B., & Wagner, K. (2019). The Inferential Language Comprehension (iLC) Framework: Supporting Children's Comprehension of Visual Narratives. *Topics in Cognitive Science*, 1–18. doi: 10.1111/tops.12457

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *Modern Language Journal*, *102*(1), 120–141. doi: 10.1111/modl.12447

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel* (Tech. Rep.). Naval Technical Training Command Millington TN Research Branch.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182. doi: 10.1037/0033-295x.95.2.163

Kintsch, W. (1998). Comprehension: A paradigm for cognition. New York:: Cambridge university press.

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

Kintsch, W., & Yarbrough, J. C. (1982). Role of rhetorical structure in text comprehension. *Journal of Educational Psychology*, *74*(6), 828–834. doi: 10.1037/0022-0663.74.6.828

Kirkorian, H. L., & Anderson, D. R. (2018). Effect of sequential video shot comprehensibility on attentional synchrony: A comparison of children and adults. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(40), 9867–9874. doi: 10.1073/pnas.1611606114

Kirsner, K. (1994). Implicit processes in second language learning. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 283–312). San Diego, CA: Academic Press.

Kitaev, N., & Klein, D. (2018). Multilingual Constituency Parsing with Self-Attention and Pre-Training. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers.* Melbourne, Vic.: Association for Computational Linguistics.

Kizilcec, R. F., Papadopoulos, K., & Sritanyaratana, L. (2014). Showing face in video instruction. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*(October), 2095–2102. doi: 10.1145/2556288.2557207

Klare, G. R. (1984). Readability. In P. D. Pearson, R. Barr, & M. L. Kamil (Eds.), *Handbook of reading research* (pp. 681–744). New York, NY: Longman.

Klare, G. R., et al. (1963). *Measurement of readability.* Iowa State University Press.

Knoeferle, P., & Guerra, E. (2016). Visually situated language comprehension. *Linguistics and Language Compass*, *10*(2), 66–82. doi: 10.1111/lnc3.12177

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the. *Human neurobiology*, *4*, 219–227.

Koehler, K., Guo, F., Zhang, S., & Eckstein, M. P. (2014). What do saliency models predict? *Journal of Vision*, *14*(3), 14–14. doi: 10.1167/14.3.14

Kostin, I. (2004). *Exploring Item Characteristics That Are Related To the Difficulty of Toefl Dialogue Items* (Vol. 2004; Tech. Rep. No. 1). Princeton. doi: 10.1002/j.2333-8504.2004.tb01938.x

Kotani, K., Ueda, S., Yoshimi, T., & Nanjo, H. (2014). A Listenability Measuring Method for an Adaptive Computer-assisted Language Learningand Teaching System. *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, 387–394.

Kotani, K., & Yoshimi, T. (2016). Measuring readability for learners of English as a foreign language by linguistic and learner features. In *Communications in computer and information science* (Vol. 593, pp. 211–222). doi: 10.1007/978-981-10-0515-2_15

Kotani, K., & Yoshimi, T. (2017). Effectiveness of linguistic and learner features for listenability measurement using a decision tree classifier. *The Journal of Information and Systems in Education*, *16*(1), 7–11.

Koyama, D., & Ockey, G. J. (2016). the Effects of Item Preview on Video-Based Multiple-Choice Listening Assessments. *Language Learning & Technology*, *20*(201), 148–165.

Kress, G. R. (2010). *Multimodality: A social semiotic approach to contemporary communication.* Taylor & Francis.

Kuperberg, G. R. (2020). Tea with milk? a hierarchical generative framework of sequential event comprehension. *Topics in Cognitive Science*.

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

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in cognitive sciences*, *12*(2), 72–79.

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12–24. doi: 10.1016/j.jslw.2016.10.003

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. doi: 10.3758/s13428-017-0924-4

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

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259–284. doi: 10.1080/01638539809545028

Lang, A., & Geiger, S. (1996). *The effects of related and unrelated cuts on television* (Vol. 20) (No. 1). doi: 10.1177/009365093020001001

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical density in L2 written production. *Applied Linguistics*, *16*(3), 307–322. doi: 10.1093/applin/16.3.307

Le Meur, O., Le Callet, P., & Barba, D. (2007). Predicting visual fixations on video based on low-level visual features. *Vision Research*, *47*(19), 2483–2498. doi: 10.1016/j.visres.2007.06.015

Le Bruyn, B., & Paquot, M. (2020). *Learner corpus research meets second language acquisition.* Cambridge University Press.

Lee, D. Y. W. (2001). Defining core vocabulary and tracking its distribution across spoken and written genres. *Journal of English Linguistics*, *29*(3), 250–278.

Lee, J.-B., & Kalva, H. (2008). *The vc-1 and h. 264 video compression standards for broadband video services* (Vol. 32). Springer Science & Business Media.

Lee, M., & Révész, A. (2020). Promoting grammatical development through captions and textual enhancement in multimodal input-based tasks. *Studies in Second Language Acquisition.*

Lee, S. S. (2001). 'A root out of a dry ground': Resolving the researcher/researched dilemma. In J. Zeni (Ed.), *Ethical issues in practitioner research* (pp. 61–71). New York: Teachers College Press.

Leech, G., & Rayson, P. (2014). *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Lesnov, R. O. (2018). Content-Rich Versus Content-Deficient Video-Based Visuals in L2 Academic Listening Tests. *International Journal of Computer-Assisted Language Learning and Teaching*, *8*(1), 15–30. doi: 10.4018/ijcallt.2018010102

Levin, D. T., & Baker, L. J. (2017). Bridging views in cinema: A review of the art and science of view integration. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(5), 1–18. doi: 10.1002/wcs.1436

Levin, D. T., & Simons, D. J. (2000). Perceiving stability in a changing world: Combining shots and intergrating views in motion pictures and the real world. *Media Psychology*, *2*(4), 357–380.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Li, C.-H. (2015). A comparative study of video presentation modes in relation to L2 listening success. *Technology, Pedagogy and Education*(June), 1–15. doi: 10.1080/1475939x.2015.1035318

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., … Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8693 Lncs*(Part 5), 740–755. doi: 10.1007/978-3-319-10602-1_48

Liss, J. M., Utianski, R., & Lansford, K. (2013). Crosslinguistic application of english-centric rhythm descriptors in motor speech disorders. *Folia Phoniatrica et Logopaedica*, *65*(1), 3–19.

Liu, H. (2007). Probability distribution of dependency distance. *Glottometrics*, *15*, 1–12.

Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, *21*, 171–193. doi: 10.1016/j.plrev.2017.03.002

Liu, N., & Han, J. (2018). A Deep Spatial Contextual Long-Term Recurrent Convolutional Network for Saliency Detection. *IEEE Transactions on Image Processing*, *27*(7), 3264–3274. doi: 10.1109/tip.2018.2817047

Liu, Y., Feng, X., & Zhou, Z. (2016). Multimodal video classification with stacked contractive autoencoders. *Signal Processing*, *120*, 761–766. doi: 10.1016/j.sigpro.2015.01.001

Lively, B. A., & Pressey, S. L. (1923). A method for measuring the vocabulary burden of textbooks. *Educational Administration and Supervision*(9), 389–398.

Londe, Z. C. (2009). The Effects of Video Media in English as a Second Language Listening Comprehension Tests. *Issues in Applied Linguistics*, *17*(1).

Lorge, I. (1944). Predicting readability. *Teachers college record*.

Loschky, L. C., Hutson, J. P., Smith, M. E., Smith, T. J., & Magliano, J. P. (2018). Viewing Static Visual Narratives through the Lens of the Scene Perception and Event Comprehension Theory (SPECT). *Empirical Comics Research: Digital, Multimodal, and Cognitive Methods*, 217–238.

Loschky, L. C., Larson, A. M., Smith, T. J., & Magliano, J. P. (2020). The scene perception & event comprehension theory (SPECT) applied to visual narratives. *Topics in Cognitive Science*, *12*(1), 311–351. doi: 10.1111/tops.12455

Louwerse, M. M., Graesser, A. C., McNamara, D. S., & Lu, S. (2009). Embodied conversational agents as conversational partners. *Applied Cognitive Psychology*, *23*(9), 1244–1255. doi: 10.1002/acp.1527

Lu, X. (2011a). A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers' Language Development. *TESOL Quarterly*, *45*(1), 36–62. doi: 10.5054/tq.2011.240859

Lu, X. (2011b). A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers' Language Development. *TESOL Quarterly*, *45*(1), 36–62. doi: 10.5054/tq.2011.240859

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *Modern Language Journal*, *96*(2), 190–208. doi: 10.1111/j.1540-4781.2011.01232_1.x

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing*, *34*(4), 493–511. doi: 10.1177/0265532217710675

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear and hearing*, *19*(1), 1–36.

Lwo, L., & Lin, M. C.-T. (2012). The effects of captions in teenagers' multimedia l2 learning. *ReCALL*, *24*(2), 188–208.

Ma, Y., & Fosler-lussier, E. (2012). Ranking-based readability assessment for early primary children ' s literature. In *Conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 548–552).

Machado, P., Romero, J., Nadal, M., Santos, A., Correia, J., & Carballal, A. (2015). Computerized measures of visual complexity. *Acta Psychologica*, *160*, 43–57. doi: 10.1016/j.actpsy.2015.06.005

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk, Vol. 2: The database* (third edit ed.; N. . Mahwah, Ed.). Mahwah, NJ: Lawrence Erlbaum.

Madan, C. R., Bayer, J., Gamer, M., Lonsdorf, T. B., & Sommer, T. (2018). Visual Complexity and Affect: Ratings Reflect More Than Meets the Eye. *Frontiers in Psychology*, *8*(January), 1–19. doi: 10.3389/fpsyg.2017.02368

Magliano, J. P., Millis, K., Ozuru, Y., & McNamara, D. S. (2007). A multidimensional framework to evaluate reading assessment tool. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 107–136). Mahwah, NJ: Lawrence Erlbaum Associates.

Magliano, J. P., Radvansky, G. A., & Copeland, D. E. (2007). Beyond language comprehension: Situation models as a form of autobiographical memory. *Higher level language processes in the brain: Inference and comprehension processes*, 379–391.

Magliano, J. P., Radvansky, G. A., Forsythe, J. C., & Copeland, D. E. (2014). Event segmentation during first-person continuous events. *Journal of Cognitive Psychology*, *26*(6), 649–661. doi: 10.1080/20445911.2014.930042

Magliano, J. P., & Zacks, J. M. (2011). The impact of continuity editing in narrative film on event segmentation. *Cognitive Science*, *35*(8), 1489–1517. doi: 10.1111/j.1551-6709.2011.01202.x

Mahapatra, D., Gilani, S. O., & Saini, M. K. (2014). Coherency Based Spatio-Temporal Saliency Detection for Video Object Segmentation. *IEEE Journal of Selected Topics in Signal Processing*, *8*(3), 454–462. doi: 10.1109/jstsp.2014.2315874

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2008). The Effects of Nonnative Accents on Listening Comprehension: Implications for ESL Assessment. *TESOL Quarterly*, *36*(2), 173. doi: 10.2307/3588329

Major, R. C., Fitzmaurice, S. M., Bunta, F., & Balasubramanian, C. (2005). Testing the effects of regional, ethnic, and international dialects of english on listening comprehension. *Language Learning*, *55*(1), 37–69. doi: 10.1111/j.0023-8333.2005.00289.x

Malvern, D., & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language* (pp. 58–71). Clevedon, UK: Multilingual Matters.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical Richness and Language Development: Quantification and Assessment.* New York: Palgrave Macmillan.

Mancas, C., & Gosseli, B. (2007). Natural scene text understanding. In *Vision systems: Segmentation and pattern recognition.* IntechOpen. doi: 10.5772/4966

Manning, C. D., Bauer, J., Finkel, J., & Bethard, S. J. (2014). The Stanford CoreNLP Natural

Language Processing Toolkit. In *Proceedings of52nd annual meeting ofthe association for computational linguistics: System demonstrations* (pp. 55–60). Baltimore, Maryland USA.

Markham, P. (1999). Captioned videotapes and second-language listening word recognition. *Foreign Language Annals*, *32*(3), 321–328.

Markham, P. L., Peter, L. A., & McCarthy, T. J. (2001). The effects of native language vs. target language captions on foreign language students' dvd video comprehension. *Foreign language annals*, *34*(5), 439–445.

Martinc, M., Pollak, S., & Robnik-Šikonja, M. (2019). Supervised and unsupervised neural approaches to text readability.

Matsuura, H., Chiba, R., Mahoney, S., & Rilling, S. (2014). Accent and speech rate effects in English as a lingua franca. *System*, *46*(1), 143–150. doi: 10.1016/j.system.2014.07.015

Matthews, J., & Cheng, J. (2015). Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System*, *52*, 1–13. doi: 10.1016/j.system.2015.04.015

Mattys, S. L., & Jusczyk, P. W. (2001). *Phonotactic cues for segmentation of fluent speech by infants* (Vol. 78) (No. 2). doi: 10.1016/s0010-0277(00)00109-8

Mayer, R. E. (2009). *Multimedia Learning*. Cambridge: Cambridge University Press. doi: 10.1017/cbo9780511811678

Mayer, R. E. (2014). Principles Based on Social Cues in Multimedia Learning: Personalization, Voice, Image, and Embodiment Principles. In R. Mayer (Ed.), *The cambridge handbook of multimedia learning* (pp. 345–368). Cambridge: Cambridge University Press. doi: 10.1017/cbo9781139547369.017

Mayer, R. E., Fiorella, L., & Stull, A. (2020). Five ways to increase the effectiveness of instructional video. *Educational Technology Research and Development*(0123456789). doi: 10.1007/s11423-020-09749-6

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)* (Unpublished PhD dissertation). University of Memphis.

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. doi: 10.3758/brm.42.2.381

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological review*, *88*(5), 375.

McDonald, S. A., & Shillcock, R. C. (2009). Rethinking the Word Frequency Effect: The Neglected Role of Distributional Information in Lexical Processing. *Language and Speech*, *44*(3), 295–322. doi: 10.1177/00238309010440030101

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlations coefficients. *Psychological Methods*, *1*(4), 390–390. doi: 10.1037/1082-989x.1.4.390

McKenna, M. C., & Layton, K. (1990). Concurrent Validity of Cloze as a Measure of Intersentential Comprehension. *Journal of Educational Psychology*, *82*(2), 372–377. doi: 10.1037/0022-0663.82.2.372

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

McNamara, D. S., Kintsch, E., Songer, N. B., Kintsch, W., Cognition, S., & Mcnamara, D. S. (1996). Are Good Texts Always Better? Interactions of Text Coherence, Background Knowledge, and Levels of Understanding in Learning from Text. *Cognition and Instruction*, *14*(1), 1–43. doi: 10.1207/s1532690xci1401

McNamara, D. S., & Kintsch, W. (1996a). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*(3), 247–288. doi: 10.1080/01638539609544975

McNamara, D. S., & Kintsch, W. (1996b). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*(3), 247–288. doi: 10.1080/01638539609544975

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing Linguistic Features of Cohesion. *Discourse Processes*, *47*(4), 292–330. doi: 10.1080/01638530902959943

Mcnamara, D. S., & Magliano, J. (2009). Toward a Comprehensive Model of Comprehension. In *Psychology of learning and motivation - advances in research and theory* (Vol. 51, pp. 297–384). doi: 10.1016/s0079-7421(09)51009-2

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.

McQueen, J. M., & Pitt, M. A. (1998). Is Compensation for Coarticulation Mediated by the Lexicon?,. *Journal of Memory and Language*, *39*(3), 347–370. doi: 10.1006/jmla.1998.2571

Meara, P., & Bell, H. (2001). P_lex: A simple and effective way of describing the lexical characteristics of short l2 texts.

Merriam-Webster. (2021a). Author.

Merriam-Webster. (2021b). Author.

Mesmer, H. A. (2008). *Tools for matching readers to texts.* Guilford Press.

Mesmer, H. A., Cunningham, J. W., & Hiebert, E. H. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading Research Quarterly*, *47*(3), 235–258. doi: 10.1002/rrq.019

Mestres, E. T., & Pellicer-Sánchez, A. (2019). Young efl learners' processing of multimodal input: Examining learners' eye movements. *System*, *80*, 212–223.

Miestamo, M., et al. (2008). Grammatical complexity in a cross-linguistic perspective. *Language complexity: Typology, contact, change*, *23*, 41.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *In advances in neural information processing systems* (pp. 3111–3119). doi: 10.1162/jmlr.2003.3.4-5.951

Miller, G. A. (1998). *Wordnet: An electronic lexical database.* MIT press.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, *3*(4), 235–244.

Milton, J. (2009). *Measuring second language vocabulary acquisition.* Bristol, UK: Multilingual Matters.

Mintz, T. H., Walker, R. L., Welday, A., & Kidd, C. (2018). Infants' sensitivity to vowel harmony and its role in segmenting speech. *Cognition*, *171*(November 2010), 95–107. doi: 10.1016/j.cognition.2017.10.020

Mirzaei, M. S., Meshgi, K., Akita, Y., & Kawahara, T. (2017). Partial and synchronized captioning: A new tool to assist learners in developing second language listening skill. *ReCALL*, *29*(2), 178–199. doi: 10.1017/s0958344017000039

Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2011). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive computation*, *3*(1), 5–24.

Mnsterberg, H. (1970). *The film: A psychological study: The silent photoplay of 1916.* New York.

Moacdieh, N., & Sarter, N. (2015). Display clutter: A review of definitions and measurement techniques. *Human Factors*, *57*(1), 61–100. doi: 10.1177/0018720814541145

Mohan, B., & Helmer, S. (1988). Context and second language development: Preschoolers' comprehension of gestures. *Applied Linguistics*, *9*(3), 275–292.

Montabone, S., & Soto, A. (2010). Human detection using a mobile platform and novel features derived from a visual saliency mechanism. *Image and Vision Computing*, *28*(3), 391–402. doi: 10.1016/j.imavis.2009.06.006

Montero Perez, M., Peters, E., & Desmet, P. (2014). Is less more? Effectiveness and

perceived usefulness of keyword and full captioned video for L2 listening comprehension. *ReCALL*, *26*(01), 21–43. doi: doi:10.1017/S0958344013000256

Montero Perez, M. (2019). Pre-learning vocabulary before viewing captioned video: An eye-tracking study. *The Language Learning Journal*, *47*(4), 460–478.

Montero Perez, M. (2020). Incidental vocabulary learning through viewing video: The role of vocabulary knowledge and working memory. *Studies in Second Language Acquisition*, 1–25.

Montero Perez, M., Peters, E., & Desmet, P. (2018). Vocabulary learning through viewing video: the effect of two enhancement techniques. *Computer Assisted Language Learning*, *31*(1-2), 1–26.

Morris, L., & Cobb, T. (2004). Vocabulary profiles as predictors of the academic performance of Teaching English as a Second Language trainees. *System*, *32*(1), 75–87. doi: 10.1016/j.system.2003.05.001

Muljani, D., Koda, K., & Moates, D. R. (1998). The development of word recognition in a second language. *Applied Psycholinguistics*, *19*, 99–113.

Nadeem, F., & Ostendorf, M. (2018). Estimating Linguistic Complexity for Science Texts. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 45–55.

Nagel, E. (1961). The structure of science.

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychological research*, *71*(1), 4–12.

Neider, M. B., & Zelinsky, G. J. (2011). Cutting through the clutter: Searching for targets in evolving complex scenes. *Journal of Vision*, *11*(14), 7–7.

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). Measures of Text Difficulty: Testing their Predictive Value for Grade Levels and Student Performance. , 58.

Newbold, N., & Gillam, L. (2010). The linguistics of readability: The next step for word processing. In *Proceedings of the naacl hlt 2010 workshop on computational linguistics and writing: Writing processes and authoring aids* (pp. 65–72).

Newtson, D., & Engquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology*, *12*(5), 436–450. doi: 10.1016/0022-1031(76)90076-7

NGACB & CCSSO. (2010). Common core state standards for english language arts and literacy in history/social studies, science, and technical subjects. *Pennsylvania Department of Education*.

Nguyen, T. V., Xu, M., Gao, G., Kankanhalli, M., Tian, Q., & Yan, S. (2013). Static saliency

vs. dynamic saliency. *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, 987–996. doi: 10.1145/2502081.2502128

Nissan, S., DeVincenzi, F., & Tang, K. L. (1995). an Analysis of Factors Affecting the Difficulty of Dialogue Items in Toefl Listening Comprehension. *ETS Research Report Series*, *1995*(2), i–42. doi: 10.1002/j.2333-8504.1995.tb01671.x

Norris, J. M., & Ortega, L. (2009). Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. *Applied Linguistics*, *30*(4), 555–578. doi: 10.1093/applin/amp044

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed ed.). New York, NY: McGraw-Hill.

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, *24*(4), 517–537. doi: 10.1177/0265532207080771

Ockey, G. J., & French, R. (2016). From One to Multiple Accents on a Test of L2 Listening Comprehension. *Applied Linguistics*, *37*(5), 693–715. doi: 10.1093/applin/amu060

Ockey, G. J., Papageorgiou, S., & French, R. (2016). Effects of Strength of Accent on an L2 Interactive Lecture Listening Comprehension Test. *International Journal of Listening*, *30*(1-2), 84–98. doi: 10.1080/10904018.2015.1056877

of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, C. (2001). *Common european framework of reference for languages: learning, teaching, assessment.* Cambridge University Press.

Ojemann, R. H. (1933). The reading difficulty of parent education materials. In *Proceedings of the iowa academy of science* (Vol. 40, pp. 159–170).

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research*, *155 B*, 23–36. doi: 10.1016/s0079-6123(06)55002-2

O'loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing*, *12*(2), 217–237. doi: 10.1177/026553229501200205

Ouwehand, K., van Gog, T., & Paas, F. (2015). Designing effective video-based modeling examples using gaze and gesture cues. *Educational Technology and Society*, *18*(4), 78–88.

Ozasa, T., Weir, G. R. S., & Fukui, M. (2014). Development of a readability index attuned to the new English Course of Study in Japan. In *Proceedings of iceri2014 conference . 7th international conference of education , research and innovation . international academy of technology , education and d.*

Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic

information from speech and gesture: Insights from event-related brain potentials. *Journal of cognitive neuroscience*, *19*(4), 605–616.

Paivio, A. (2007). Mind and its evolution: A dual coding theoretical interpretation. Mahwah, NJ: Lawrence erlbaum Associates, Inc.

Paivio, A., Walsh, M., & Bons, T. (1994). Concreteness effects on memory: When and why? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1196–1204. doi: 10.1037/0278-7393.20.5.1196

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*(1), 117–134. doi: 10.1177/0267658314536435

Palumbo, L., Makin, A. D. J., & Bertamini, M. (2014). Examining visual complexity and its influence on perceived duration. *Journal of Vision*, *14*(14), 1–18. doi: 10.1167/14.14.3.doi

Pardo-Ballester, C. (2016). Using Video in Web-Based Listening Tests. *Journal of New Approaches in Educational Research*, *5*(2), 91–98. doi: 10.7821/naer.2016.7.170

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International conference on machine learning* (pp. 1310–1318).

Pashler, H. E. (1999). *The psychology of attention.* MIT press.

Paterson, D. G., & Tinker, M. A. (1940). *How to make type readable.* Harper.

Pattemore, A., & Muñoz, C. (2020). Learning l2 constructions from captioned audio-visual exposure: The effect of learner-related factors. *System*, *93*, 102303.

Pearson, P. D., & Hiebert, E. H. (2014). The state of the field: Qualitative analyses of text complexity. *The Elementary School Journal*, *115*(2), 161–183. doi: 10.1086/521238

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. doi: 10.1007/s13398-014-0173-7.2

Perez, M. M., & Rodgers, M. P. H. (2019). Video and language learning. *The Language Learning Journal*, *47*(4), 403–406. doi: 10.1080/09571736.2019.1629099

Perfetti, C., & Stafura, J. (2014). Word Knowledge in a Theory of Reading Comprehension. *Scientific Studies of Reading*, *18*(1), 22–37. doi: 10.1080/10888438.2013.827687

Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *TESOL Quarterly*, *53*(4), 1008–1032.

Peters, E., & Muñoz, C. (2020). Introduction to special issue language learning from multimodal input. *Studies In Second Language Acquisition.*

Petersen, S. E., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech and Language*, *23*(1), 89–106. doi: 10.1016/j.csl.2008.04.003

Peterson, M. (2006). Learner interaction management in an avatar and chat-based virtual world. , *19*(1), 79–103. doi: 10.1080/09588220600804087

Pieters, R., Wedel, M., & Batra, R. (2010). The Stopping Power of Advertising: Measures and Effects of Visual Complexity. *Journal of Marketing*, *74*(5), 48–60. doi: 10.1509/jmkg.74.5.48

Pilán, I. (2018). *Automatic proficiency level prediction for Intelligent Computer-Assisted Language Learning* (Unpublished doctoral dissertation). University of Gothenburg.

Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. *Proceedings of the Conference on Empirical ...*(October), 186–195. doi: anthology/D08-1020

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing : From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98–125. doi: 10.1016/j.inffus.2017.02.003

Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of experimental psychology: General*, *109*(2), 160.

Potter, R. F., Jamison-Koenig, E. J., Lynch, T., & Sites, J. (2016). Effect of Vocal-Pitch Difference on Automatic Attention to Voice Changes in Audio Messages. *Communication Research*, 1–18. doi: 10.1177/0093650215623835

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *the Journal of the Acoustical Society of America*, *90*(6), 2956–2970.

Puimège, E., & Peters, E. (2019). Learning L2 vocabulary from audiovisual input: an exploratory study into incidental learning of single words and formulaic sequences. *Language Learning Journal*, *1736*. doi: 10.1080/09571736.2019.1638630

Puimège, E., & Peters, E. (2020). Learning Formulaic Sequences Through Viewing L2 Television and Factors That Affect Learning. *Studies in Second Language Acquisition*, 1–25. doi: 10.1017/s027226311900055x

Pujadas, G., & Muñoz, C. (n.d.). Examining adolescent efl learners' tv viewing comprehension through captions and subtitles. *Studies in Second Language Acquisition*, 1–25.

Pujadas, G., & Muñoz, C. (2019). Extensive viewing of captioned and subtitled tv series: A study of l2 vocabulary learning by adolescents. *The Language Learning Journal*, *47*(4), 479–496.

Pujolă, J.-T. (2002). Calling for help: Researching language learning strategies using help facilities in a web-based multimedia program. *ReCALL*, *14*(2), 235–262.

Pyle, D. (1999). *Data Preparation for Data Mining*. morgan kaufmann.

Qi, P., Dozat, T., Zhang, Y., & Manning, C. D. (2019). Universal Dependency Parsing from Scratch. In *Proceedings of the conll 2018 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 160–170).

Radvansky, G. A., & Zacks, J. M. (2014). *Event Cognition* (Vol. 6) (No. 38). Oxford University Press. doi: 10.1093/acprof:oso/9780199898138.001.0001

Radvansky, G. A., & Zacks, J. M. (2017). Event boundaries in memory and cognition. *Current Opinion in Behavioral Sciences*, *17*, 133–140. doi: 10.1016/j.cobeha.2017.08.006

Rayner, K., & Reichle, E. D. (2010). Models of the reading process. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 787–799.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Redish, J. (2000). Readability formulas have even more limitations than Klare discusses. *ACM Journal of Computer Documentation*, *24*(3), 132–137. doi: 10.1145/344599.344637

Reeves, B., Thorson, E., Rothschild, M. L., McDonald, D., Hirsch, J., & Goldstein, R. (1985). Attention to television: Intrastimulus effects of movement and scene changes on alpha variation over time. *International Journal of Neuroscience*, *27*(3-4), 241–255.

Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., & Gajos, K. Z. (2013). Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, 2049. doi: 10.1145/2470654.2481281

Renandya, W. A., & Farrell, T. S. (2011). 'Teacher, the tape is too fast!' Extensive listening in ELT. *ELT Journal*, *65*(1), 52–59. doi: 10.1093/elt/ccq015

Rescher, N. (1998). *Complexity: A philosophical overview*. Transaction Publishers.

Révész, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, *35*(01), 31–65. doi: 10.1017/s0272263112000678

Révész, A., Kourtali, N. E., & Mazgutova, D. (2017). Effects of Task Complexity on L2 Writing Behaviors and Linguistic Complexity. *Language Learning*, *67*(1), 208–241. doi: 10.1111/lang.12205

Richard, J. C., Platt, J., & Platt, H. (1992). Dictionary of language teaching & applied linguistics. *Essex: Longman*.

Rigau, J., Feixas, M., & Sbert, M. (2005). An information-theoretic framework for image complexity. In *In proceedings of the first eurographics conference on computational aesthetics in graphics, visualization and imaging* (pp. 177–184).

Rodgers, M. P., & Webb, S. (2017). The effects of captions on EFL learners' comprehension

of english-language television programs. *CALICO Journal*, *34*(1), 20–38. doi: 10 .1558/cj.29522

Rodgers, M. P., & Webb, S. (2020). Incidental vocabulary learning through viewing television. *ITL-International Journal of Applied Linguistics*.

Roncaglia-Denissen, M. P., Schmidt-Kassow, M., & Kotz, S. A. (2013). Speech Rhythm Facilitates Syntactic Ambiguity Resolution: ERP Evidence. *PLoS ONE*, *8*(2), 1–10. doi: 10.1371/journal.pone.0056000

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382–439.

Rosenholtz, R., Li, Y., Mansfield, J., & Jin, Z. (2005). Feature congestion: A Measure of Display Clutter. In *Proceedings of the sigchi conference on human factors in computing systems - chi '05* (p. 761). doi: 10.1145/1054972.1055078

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of vision*, *7*(2), 17.1–22. doi: 10.1167/7.2.17

Rost, M. (2013). *Teaching and Researching Listening* (2d ed., Vol. 38) (No. 4). Ny: Routledge. doi: 10.1016/j.dss.2003.08.004

Roy, C. M., Susan, M. F., Ferenc, B., & Chandrika, B. (2005). Testing the Effects of Regional, Ethnic, and International Dialects of English on Listening Comprehension. *Language Learning*, *55*(1), 37–69. doi: 10.1111/j.0023-8333.2005.00289.x

RStudio Team. (2016). *RStudio: Integrated Development for R.* Boston, MA: RStudio, Inc.

Rubin, J. (1994). A review of language listening comprehension research. *The Modern Language Journal*, *78*(2), 199–221.

Rubin, J. (1995). The contribution of video to the development of competence in listening. In D. J. Mendelsohn & J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 151–165). San Diego: Dominie Press.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Ii. the contextual enhancement effect and some tests and extensions of the model. *Psychological review*, *89*(1), 60.

Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing*, *1*(3-4), 185–216. doi: 10.1080/15305058 .2001.9669470

Ryan, J., & Granville, S. (2020). The suitability of film for modelling the pragmatics of interaction: Exploring authenticity. *System*, *89*, 102186.

Saito, K. (2013). Age effects on late bilingualism: The production development of /[U+0279]/ by high-proficiency japanese learners of english. *Journal of Memory*

*and Language*, *69*(4), 546–562.

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical Profiles of Comprehensible Second Language Speech. *Studies in Second Language Acquisition*, *38*(04), 677–701. doi: 10.1017/s0272263115000297

Salomon, D. (2004). *Data compression: the complete reference*. Springer Science & Business Media.

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, *27*(3), 343–360. doi: 10.1177/0267658310395851

Sanders, T. J., Spooren, W. P., & Noordman, L. G. (1992). Toward a taxonomy of coherence relations. *Discourse processes*, *15*(1), 1–35.

Satapathy, S. C., Govardhan, A., Raju, K. S., & Mandal, J. K. (2015). Video shot boundary detection: A review. *Advances in Intelligent Systems and Computing*, *338*, I–iv. doi: 10.1007/978-3-319-13731-5

Sawyer, M. H. (1991). A review of research in revising instructional text. *Journal of Reading Behavior*, *23*(3), 307–333.

Sayood, K. (2017). *Introduction to data compression*. Morgan Kaufmann.

Schneider, S., Beege, M., Nebel, S., & Rey, G. D. (2018). A meta-analysis of how signaling affects learning with media. *Educational Research Review*, *23*(August 2017), 1–24. doi: 10.1016/j.edurev.2017.11.001

Schnotz, W. (2013). No Title. , 1–49.

Schomaker, J., Walper, D., Wittmann, B. C., & Einhäuser, W. (2017). Attention in natural scenes: Affective-motivational factors guide gaze independently of visual salience. *Vision Research*, *133*, 161–175. doi: 10.1016/j.visres.2017.02.003

Schwan, S., Garsoffky, B., & Hesse, F. W. (2000). Do film cuts faciliate the perceptual and cognitive organisation of activity sequences? *Memory and Cognition*, *28*(2), 214–223.

Schwan, S., & Ildirar, S. (2010). Watching film for the first time: How adult viewers interpret perceptual discontinuities in film. *Psychological Science*, *21*(7), 970–976. doi: 10.1177/0956797610372632

Schwanenflugel, P. (1991). Why are abstract concepts hard to understand? In P. Schwanenflugel (Ed.), *The psychology of word meanings* (p. 223–250). Hillsdale, NJ: Erlbaum.

Schwarm, S. E., & Ostendorf, M. (2005). Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *Proceedings of the 43rd annual meeting of the acl* (pp. 523–530).

Sehairi, K., Chouireb, F., & Meunier, J. (2017). Comparative study of motion detection methods for video surveillance systems. *Journal of Electronic Imaging*, *26*(2), 023025. doi: 10.1117/1.jei.26.2.023025

Seo, K. (2005). Development of a listening strategy intervention program for adult learners of japanese. *International Journal of Listening*, *19*(1), 63–78.

Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly*, *17*(2), 229–255. doi: 10.1037/0022-0663.82.2.372

Sheehan, K. M. (2017). Validating Automated Measures of Text Complexity. *Educational Measurement: Issues and Practice*, *00*(0), 1–9. doi: 10.1111/emip.12155

Sheehan, K. M., Flor, M., & Napolitano, D. (2013). A Two-Stage Approach for Generating Unbiased Estimates of Text Complexity. In *Proceedings of the workshop on natural language processing for improving textual accessibility* (pp. 49–58). doi: 10.1016/j.biopsych.2012.07.028

Sheehan, K. M., Kostin, I., & Futagi, Y. (2010). *Generating Automated Text Complexity Classifications That Are Aligned With Targeted Text Complexity Standards* (Tech. Rep. No. December). doi: 10.1002/j.2333-8504.2010.tb02235.x

Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The TextEvaluator tool. *Elementary School Journal*, *115*(2), 184–209. doi: 10.1086/678294

Sherman, A. L. (1893). *Analytics of literature: A manual for the objective study of english prose and poetry*. Boston: Ginn Co.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. doi: 10.1214/10-sts330

Si, L., & Callan, J. (2001). A statistical model for scientific readability. In *Proceedings of the tenth international conference on information and knowledge management* (pp. 574–576).

Sidaty, N., Larabi, M. C., & Saadane, A. (2017). Toward an audiovisual attention model for multimodal video content. *Neurocomputing*, *259*, 94–111. doi: 10.1016/j.neucom.2016.08.130

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487–512. doi: 10.1093/applin/amp058

Singh, L., White, K. S., & Morgan, J. L. (2008). Building a Word-Form Lexicon in the Face of Variable Input: Influences of Pitch and Amplitude on Early Spoken Word Recognition. *Language Learning and Development*, *4*(2), 157–178. doi: 10.1080/15475440801922131

Smidt, E., & Hegelheimer, V. (2004). *Effects of online academic lectures on ELS listening comprehension, incidental vocabulary acquisition, and strategy use* (Vol. 17) (No. 5). Routledge, part of the Taylor & Francis Group. doi: 10.1080/0958822042000319692

Smith, L. E., & Bisazza, J. A. (1978). The comprehensiblity of English for college students in seven countries. *Language Learning*, *32*(2), 259–269.

Smith, L. E., & Bisazza, J. A. (1982). The comprehensibility of three varieties of English for college students in seven countries. *Language Learning*, *32*(2), 259–269. doi: 10.1111/lang.1982.32.issue-2

Snodgrass, J. G., & Vanderwart, M. (1980). A standardised set of 260 pictures: normal for name agreement, familiarity and visual complexity. *Journal of Experimental Psychology: General*, *6*(2), 174–215.

Spanjers, I. A., van Gog, T., & van Merriënboer, J. J. (2010). A Theoretical Analysis of How Segmentation of Dynamic Visualizations Optimizes Students' Learning. *Educational Psychology Review*, *22*(4), 411–423. doi: 10.1007/s10648-010-9135-6

Srivastava, N., Velloso, E., Lodge, J. M., Erfani, S., & Bailey, J. (2019). Continuous evaluation of video lectures from real-time difficulty self-report. *Conference on Human Factors in Computing Systems - Proceedings*. doi: 10.1145/3290605.3300816

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign languag. *Studies in Second Language Acquisition*, *31*(4), 577–607. doi: 10.1017/s0272263109990039

Stenner, A. J., Horabin, I., Smith, D. R., & Smith, M. (1988). Most comprehension test do measure reading comprehension: A response to mclean and goldstein. *Phi Delta Kappan*, 765–767.

Stenner, J. A. (1996). Measuring Reading Comprehension with the Lexile Framework. In *The north american conference on adolescent/adult literacy* (pp. 1–29).

Strain, E., & Herdman, C. M. (1999). Imageability effects in word naming: An individual differences analysis. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *53*(4), 347.

Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1140.

Stull, A. T., Fiorella, L., & Mayer, R. E. (2018). An eye-tracking analysis of instructor presence in video lectures. *Computers in Human Behavior*, *88*(July), 263–272. doi: 10.1016/j.chb.2018.07.019

Suárez, M. d. M., & Gesa, F. (2019). Learning vocabulary with the support of sustained exposure to captioned video: do proficiency and aptitude make a difference? *The*

*Language Learning Journal*, *47*(4), 497–517.

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*(4), 661–699. doi: 10 .1111/j.0023-8333.2005.00320.x

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, *26*(2), 212–215.

Sung, Y. T., Lin, W. C., Dyson, S. B., Chang, K. E., & Chen, Y. C. (2015). Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *Modern Language Journal*, *99*(2), 371–391. doi: 10.1111/modl.12213

Suvorov, R. (2009). Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format. *Developing and evaluating language learning materials*, 53–68.

Suvorov, R. (2014a). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing*, 1–21. doi: 10.1177/0265532214562099

Suvorov, R. (2014b). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing*, 1–21. doi: 10.117702655322

Suzuki, Y., Nakata, T., & Dekeyser, R. (2019). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *Modern Language Journal*, *103*(3), 713–720. doi: 10.1111/modl.12585

Tanaka, J., Weiskopf, D., & Williams, P. (2001). The role of color in high-level vision. *Trends in Cognitive Sciences*, *5*(5), 211–215. doi: 10.1016/s1364-6613(00)01626-0

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Taylor, G. (2005). Perceived processing strategies of students watching captioned video. *Foreign Language Annals*, *38*(3), 422–427. doi: 10.1111/j.1944-9720.2005.tb02228.x

Taylor, W. L. (1953). "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, *30*(4), 415–433.

Templin, M. (1957). *Certain language skills in children: Their development and interrelationships*. Minneapolis: The University of Minnesota Press.

Teng, F. (2019). Maximizing the potential of captions for primary school esl students' comprehension of english-language videos. *Computer Assisted Language Learning*, *32*(7), 665–691.

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & psychophysics*, *51*(6), 599–606.

Theeuwes, J. (2010). Top-down and bottom-up control of visual selection. *Acta Psychologica*, *135*(2), 77–99. doi: 10.1016/j.actpsy.2010.02.006

Thomas, F., & Cédrick, F. (2012). An " AI readability " formula for French as a foreign language. *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*,(July), 466–477.

Thompson, S. E. (2003). Text-structuring metadiscourse, intonation and the signalling of organisation in academic lectures. *Journal of English for Academic Purposes*, *2*(1), 5–20. doi: 10.1016/s1475-1585(02)00036-x

Thorndike, E. L. (1921). The teacher's word book.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms.* Hillsdale, NJ: Lawrence Erlbaum.

Tomar, S. (2006). Converting video formats with FFmpeg. *Linux Journal*, *2006*(146), 10.

Toyama, Y., Hiebert, E. H., & Pearson, P. D. (2017). An Analysis of the text complexity of leveled passages in four popular classroom reading assessments. *Educational Assessment*, *22*(3), 139–170. doi: 10.1080/10627197.2017.1344091

Trabasso, T., & Nickels, M. (1992). The development of goal plans of action in the narration of a picture story. *Discourse processes*, *15*(3), 249–275.

Tree, J. E. (1995). *The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech* (Vol. 34) (No. 6). doi: 10.1006/jmla.1995.1032

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

Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to basics: How Measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, *39*(3), 302–327. doi: 10.1093/applin/amw009

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. doi: 10.1016/0010-0285(80)90005-5

Tsao, Y. C., Weismer, G., & Iqbal, K. (2006). Interspeaker variation in habitual speaking rate: Additional evidence. *Journal of Speech, Language, and Hearing Research*, *49*(5), 1156–1164. doi: 10.1044/1092-4388(2006/083)

Tsay, R. S. (2013). *Multivariate time series - R and financial applications.* John Wiley & Sons.

Tuch, A. N., Bargas-Avila, J. A., Opwis, K., & Wilhelm, F. H. (2009). Visual complexity of websites: Effects on users' experience, physiology, performance, and memory. *International journal of human-computer studies*, *67*(9), 703–715.

Tuinman, A., Mitterer, H., & Cutler, A. (2012). Resolving ambiguity in familiar and unfamiliar casual speech. *Journal of Memory and Language*, *66*(4), 530–544. doi: 10.1016/j.jml.2012.02.001

Tuldava, J. (1996). The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics*, *3*(1), 38–50. doi: 10.1080/09296179608590062

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures in lexical richness in perspective. *Computers and the Humanities*, *32*(5), 323–352. doi: 10.1023/a:1001749303137

Ure, J. (1971a). Lexical density: A computational technique and some findings. In M. Coultard (Ed.), *Talking about text* (pp. 27–48). Birmingham, England: English Language Research, University of Birmingham.

Ure, J. (1971b). Lexical density and register differentiation. In G. E. Perren & J. L. M. Trim (Eds.), *Applications of linguistics: Selected papers of the second international congress of applied linguistics, cambridge 1969.* Cambridge.

Vaden, K., Halpin, H., & Hickok, G. (2009). *Irvine Phonotactic Online Dictionary, Version 2.0. [Data file].*

Vajjala, S., & Lucic, I. (2018). *OneStopEnglish corpus : A new corpus for automatic readability assessment and text simplification.*

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *In proceedings of the seventh workshop on building educational applications using nlp* (pp. 163–173).

Vajjala, S., & Meurers, D. (2014). Exploring measures of "readability" for spoken language: Analyzing linguistic features of subtitles to identify age-specific TV programs. In *Procedings of the 3rd workshop on predicting and improving text readability for target reader populations (pitr)* (pp. 21–29).

Van Gog, T., Verveer, I., & Verveer, L. (2014). Learning from video modeling examples: Effects of seeing the human model's face. *Computers and Education*, *72*, 323–327. doi: 10.1016/j.compedu.2013.12.004

Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*(4), 457–479. doi: 10.1093/applin/ams074

van den Broek, P., Risden, K., Fletcher, C. R., & Thurlssssow, R. (1996). A "landscape" view of reading: Fluctuating patterns of activation and the construction of a stable memory representation. In *Models of understanding text* (pp. 165–187).

Vandergrift, L. (2004). Listening to learn or learning to listen? *Annual review of applied linguistics*, *24*, 3.

Vandergrift, L., & Cross, J. (2014). Captioned video: How much listening is really going on? *Contact*(40), 31–3.

Vandergrift, L., & Goh, C. C. (2012). *Teaching and learning second language listening: Metacognition in action*. Routledge.

Vanderplank, R. (2010). Déj vu A decade of research on language laboratories, television and video in language learning. *Language Teaching*, *43*(1), 1–37. doi: 10.1017/s0261444809990267

Vanderplank, R. (2019). 'gist watching can only take you so far': attitudes, strategies and changes in behaviour in watching films with captions. *The Language Learning Journal*, *47*(4), 407–423.

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., … Yu, T. (2014). Scikit-image: image processing in Python. *PeerJ*, *2*, e453. doi: 10.7717/peerj.453

van Dijk, T. A., & Kintsch, W. (1983). *Strategies ofdiscourse comprehension*. New York: Academic Press.

Van Leeuwen, T. (2004). Ten reasons why linguists should pay attention to visual communication. *Discourse and technology: Multimodal discourse analysis*, 7–19.

van Wermeskerken, M., Ravensbergen, S., & van Gog, T. (2017). Effects of instructor presence in video modeling examples on attention and learning. *Computers in Human Behavior*, 1–9. doi: 10.1016/j.chb.2017.11.038

van Wermeskerken, M., & van Gog, T. (2017). Seeing the instructor's face and gaze in demonstration video examples affects attention allocation but not learning. *Computers and Education*, *113*, 98–107. doi: 10.1016/j.compedu.2017.05.013

Vatterott, D. B., & Vecera, S. P. (2012). Experience-dependent attentional tuning of distractor rejection. *Psychonomic Bulletin and Review*, *19*(5), 871–878. doi: 10.3758/s13423-012-0280-4

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Processes*, *56*(3), 229–255. doi: 10.1080/0163853x.2018.1448677

Vilà-Giménez, I., Igualada, A., & Prieto, P. (2019). Observing storytellers who use rhythmic beat gestures improves children's narrative discourse performance. *Developmental Psychology*, *55*(2), 250.

Vitevitch, M. (2007). The spread of the phonological neighborhood influences spoken word recognition. *Memory & Cognition*, *35*(1), 166–175. doi: 10.3758/bf03195952

Vitevitch, M., & Luce, P. (2016). Phonological Neighborhood Effects in Spoken Word Perception and Production. *Ssrn*. doi: 10.1146/annurev-linguistics-030514-124832

Vitevitch, M., & Luce, P. A. (2004). A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 481–487. doi: 10.3758/bf03195594

Vogel, M., & Washburne, C. (1928). An objective method of determining grade placement of children's reading material. *The Elementary School Journal*, *28*(5), 373–381.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard.

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology*, *11*(1), 67–86.

Wagner, E. (2008). Video Listening Tests: What Are They Measuring? *Language Assessment Quarterly*, *5*(3), 218–243. doi: 10.1080/15434300802213015

Wagner, E. (2010a). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, *27*(4), 493–513. doi: 10.1177/0265532209355668

Wagner, E. (2010b). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, *27*, 493–513. doi: 10.1177/0265532209355668

Wagner, E. (2010c). Test-takers' interaction with an L2 video listening test. *System*, *38*(2), 280–291. doi: 10.1016/j.system.2010.01.003

Wagner, E. (2013). An Investigation of How the Channel of Input and Access to Test Questions Affect L2 Listening Test Performance. *Language Assessment Quarterly*, *10*(2), 178–195. doi: 10.1080/15434303.2013.769552

Wagner, E. (2014). Using Unscripted Spoken Texts in the Teaching of Second Language Listening. *TESOL Journal*, *5*(2), 288–311. doi: 10.1002/tesj.120

Wang, B., & Dudek, P. (2014). A fast self-tuning background subtraction algorithm. In *Ieee computer society conference on computer vision and pattern recognition workshops* (pp. 401–404). doi: 10.1109/cvprw.2014.64

Wang, J., & Antonenko, P. D. (2017). Instructor presence in instructional video: Effects on visual attention, recall, and perceived learning. *Computers in Human Behavior*, *71*, 79–89. doi: 10.1016/j.chb.2017.01.049

Wang, W., Shen, J., & Shao, L. (2017). Video Salient Object Detection via Fully Convolutional Networks. *System*, *65*, 1–12. doi: 10.1109/tip.2017.2754941

Wang, Y. T. (2019). Effects of l1/l2 captioned tv programs on students' vocabulary learning and comprehension. *CALICO Journal*, *36*(3).

Wang, Z., Bovik, A., Sheikh, H., & Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. doi: 10.1109/tip.2003.819861

Waples, D., & Tyler, R. W. (1931). *What people want to read abouy.* Chicago: University of Chicago Press.

Watanabe, M., Hirose, K., Den, Y., & Minematsu, N. (2008). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication*, *50*(2), 81–94. doi: 10.1016/j.specom.2007.06.002

Weil, S. A. (2001). *Foreign accented speech: Adaptation and generalization* (Master's thesis). Ohio State University.

Weyers, J. R. (1999). The Effect of Authentic Video on Communicative Competence. *The Modern Language Journal*, *83*(3), 339–349. doi: 10.1111/0026-7902.00026

Whaley, C. P. (1978). Word—nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, *17*(2), 143–154.

Wilberschied, L., & Berman, P. M. (2004). Effect of Using Photos from Authentic Video as Advance Organizers on Listening Comprehension in an FLES Chinese Class. *Foreign Language Annals*, *37*(4), 534. doi: 10.1111/j.1944-9720.2004.tb02420.x

Williams, H., & Thorne, D. (2000). The value of teletext subtitling as a medium for language learning. *System*, *28*(2), 217–228.

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, *20*(1), 6–10. doi: 10 .3758/bf03202594

Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology*, *14*(1), 65–86.

Winke, P., Gass, S., & Sydorenko, T. (2013). Factors influencing the use of captions by foreign language learners: An eye[U+2010]tracking study. *The Modern Language Journal*, *97*(1), 254–275. doi: 10.1111/j.1540-4781.2012.01432.x

Winke, P., Isbell, D. R., & Ahn, J. (2019). How captions help people learn languages: A working-memory, eye-tracking study. *Language Learning & Technology*, *23*(2), 84–104.

Winkler, S. (2012). Analysis of public image and video databases for quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, *6*(6), 616–625.

Winslett, G. (2014). What Counts as Educational Video?: Working toward Best Practice Alignment between Video Production Approaches and Outcomes. *Australasian Journal of Educational Technology*, *30*(5), 487–502. doi: https://doi.org/10.14742/ajet.458

Wisniewska, N., & Mora, J. C. (2020). Can captioned video benefit second language pronunciation? *Studies in Second Language Acquisition*, 1–26.

Witten, I. H., Frank, E., Hall, M., & Pal, C. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Witzel, C., & Gegenfurtner, K. R. (2018). Color Perception: Objects, Constancy, and Categories. *Annual Review of Vision Science*, *4*(1), 475–499. doi: 10.1146/annurev-vision-091517-034231

Wolfe, J. M. (1994a). Guided Search 2 . 0 A revised model of visual search. *Psychnomic Bulletin & Review*, *1*(2), 202–238. doi: 10.3758/bf03200774

Wolfe, J. M. (1994b). Visual search in continuous, naturalistic stimuli. *Vision research*, *34*(9), 1187–1195.

Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. In *Integrated models of cognitive systems* (pp. 99–119). Oxford University Press. doi: 10.1093/acprof:oso/9780195189193.003.0008

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, *5*(6), 495–501. doi: 10.1038/nrn1411

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, *1*(3), 1–8. doi: 10.1038/s41562-017-0058

Wolfe, J. M., & Utochkin, I. S. (2019). What is a preattentive feature? *Current Opinion in Psychology*, *29*, 19–26. doi: 10.1016/j.copsyc.2018.11.005

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in cognitive sciences*, *15*(2), 77–84.

Wong, K. M., & Samudra, P. G. (2019). L2 vocabulary learning from educational media: extending dual-coding theory to dual-language learners. *Computer Assisted Language Learning*, 1–23.

Wong, S., Tsui, J., Chow, B. W. Y., Leung, V. W. H., Mok, P., & Chung, K. K. H. (2017). Perception of Native English Reduced Forms in Adverse Environments by Chinese Undergraduate Students. *Journal of Psycholinguistic Research*, *46*(5), 1–17. doi: 10

.1007/s10936-017-9486-y

Wong, S. W., Mok, P. P., Chung, K. K. H., Leung, V. W., Bishop, D. V., & Chow, B. W. Y. (2017). Perception of Native English Reduced Forms in Chinese Learners: Its Role in Listening Comprehension and Its Phonological Correlates. *TESOL Quarterly*, *51*(1), 7–31. doi: 10.1002/tesq.273

Wong, S. W. L., Mok, P. P. K., Chung, K. K.-H. H., Leung, V. W. H., Bishop, D. V. M., & Chow, B. W.-Y. Y. (2015). Perception of Native English Reduced Forms in Chinese Learners: Its Role in Listening Comprehension and Its Phonological Correlates. *TESOL Quarterly*, *0*(0), 1–25. doi: 10.1002/tesq.273

Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics*, *21*(4), 463–489. doi: 10.1093/applin/21.4.463

Wu, T. (1993). An accurate computation of the hypergeometric distribution function. *ACM Transactions on Mathematical Software*, *19*(1), 33–43. doi: 10.1145/151271.151274

Xia, M., Kochmar, E., Briscoe, T., & Building, W. G. (2016). Text Readability Assessment for Second Language Learners. In *Proceedings ofthe 11th workshop on innovative use ofnlp for building educational applications* (pp. 12–22).

Xia Wu. (2016). *Tertiary education: Issues and perspectives from Asian contexts.*

Xu, J., & Yue, S. (2014). Mimicking visual searching with integrated top down cues and low-level features. *Neurocomputing*, *133*, 1–17. doi: 10.1016/j.neucom.2013.11.037

Xu, Y., Dong, J., Zhang, B., & Xu, D. (2016). Background modeling methods in video analysis: A review and comparative evaluation. *CAAI Transactions on Intelligence Technology*, *1*(1), 43–60. doi: 10.1016/j.trit.2016.03.005

Yanagawa, K., & Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System*, *36*(1), 107–122. doi: 10.1016/j.system.2007.12.003

Yang, H.-Y. (2014). The Effects of Advance Organizers and Subtitles on EFL Learners' Listening Comprehension Skills. *CALICO Journal*, *31*(3), 345–373. doi: 10.11139/cj.31.3.345-373

Yang, J. C., & Chang, P. (2014). Captions and reduced forms instruction: The impact on EFL students' listening comprehension. *ReCALL*, *26*(01), 44–61. doi: 10.1017/s0958344013000219

Yao, Y. (2011). *The Effects of Phonological Neighborhoods on Pronunciation Variation in Conversational Speech* (Unpublished doctoral dissertation). University of California, Berkeley.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology:

Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. doi: 10.1177/1745691617693393

Yeldham, M. (2018). *Viewing L2 captioned videos: what's in it for the listener?* (Vol. 8221). Taylor & Francis. doi: 10.1080/09588221.2017.1406956

Ying-hui, H. (2006). An investigation into the task features affecting EFL listening comprehension test performance. *Asian EFL Journal*, *8*(4), 33–54.

Yoon, H. J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System*, *66*, 130–141. doi: 10.1016/j.system.2017.03.007

Yoon, S., Cho, Y., & Napolitano, D. (2016). Spoken text difficulty estimation using linguistic features. In *Proceedings of the 11th workshop on innovative use of nlp for building educational applications,* (pp. 267–276).

Young Suh, S., & van den Broek, P. (1989). Logical Necessity and Transitivity of Causal Relations in Stories. *Discourse Processes*, *12*(1), 1–25. doi: 10.1080/01638538909544717

*Youtube in numbers.* (2021). YouTube.

Yu, H., & Winkler, S. (2013). Image complexity and spatial information. *2013 5th International Workshop on Quality of Multimedia Experience, QoMEX 2013 - Proceedings*, 12–17. doi: 10.1109/QoMEX.2013.6603194

Zabrucky, K. M., & Moore, D. W. (1999). Influence of text genre on adults' monitoring of understanding and recall. *Educational Gerontology*, *25*(8), 691–710. doi: 10.1080/036012799267440

Zacks, J. M. (2003). Film, Narrative, and Cognitive Neuroscience. *Art and the Senses, ed. DP Melcher and F. Bacci.(New*, *1*, 1–20.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, *133*(2), 273–293. doi: 10.1037/0033-2909.133.2.273

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conceptions. *Psychological Bulletin*, *127*, 3–21.

Zakaluk, B., & Samuels, S. (1988). *Readability: its past, present and future.*

Zampieri, M., Malmasi, S., Paetzold, G., & Specia, L. (2017). Complex Word Identification: Challenges in Data Annotation and System Performance.

Zareva, A. (2007). Structure of the second language mental lexicon: how does it compare to native speakers' lexical organization? *Second language research*, *23*(2), 123–153.

Zhao, Q., & Koch, C. (2013). Learning saliency-based visual attention: A review. *Signal Processing*, *93*(6), 1401–1407. doi: 10.1016/j.sigpro.2012.06.014

Zhao, Y. (1997). The Effects of Listeners' Control of Speech Rate on Second Language Comprehension. *Applied Linguistics*, *18*(1), 49–68. doi: 10.1093/applin/18.1.49

Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. (2017). EAST: An efficient and accurate scene text detector. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-Janua*, 2642–2651. doi: 10.1109/cvpr.2017.283

Zhou, Z. (2012). *Ensemble methods: Foundations and algorithms*. Chapman and Hall/CRC. doi: 10.1201/b12207

Zipf, G. (1945). The meaning-frequency relationship of words. *J. Gen. Psychol.*, *33*, 251–256.

Zivkovic, Z., & Van Der Heijden, F. (2006). Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, *27*(7), 773–780. doi: 10.1016/j.patrec.2005.11.005

Zwaan, R. A. R., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological bulletin*, *123*(2), 162–185. doi: 10.1037/0033-2909.123.2.162