The Relative Contribution of Language Complexity to Second Language Video Lectures Difficulty Assessment

EMAD A. ALGHAMDI,[1] PAUL GRUBA,[2] AND EDUARDO VELLOSO[3]
[1] *King Abdulaziz University, English Language Institute, Jeddah, 21589, Saudi Arabia E-mail: eaalghamdi@kau.edu.sa*
[2] *The University of Melbourne, Department of Linguistics and Applied Linguistics, Grattan Street, Parkville, Melbourne, Victoria, 3010, Australia E-mail: p.gruba@unimelb.edu.au*
[3] *The University of Melbourne, School of Computing and Information Systems, Grattan Street, Parkville, Melbourne, Victoria, 3010, Australia E-mail: Eduardo.velloso@unimelb.edu.au*

Although core in the teaching of academic language skills, little research to date has investigated what makes video-recorded lectures difficult for language learners. As part of a larger program to develop automated videotext complexity measures, this study reports on selected dimensions of linguistic complexity to understand how they contribute to overall videotext difficulty. Based on the ratings of English language learners of 320 video lectures, we built regression models to predict subjective estimates of video lecture difficulty. The results of our analysis demonstrate that a 4-component partial least square regression model explains 52% of the variance in video difficulty and significantly outperformed a baseline model in predicting the difficulty of videos in an out-of-sample testing set. The results of our study point to the use of linguistic complexity features for predicting overall videotext difficulty and raise the possibility of developing automated systems for measuring video difficulty, akin to those already available for estimating the readability of written materials.

Keywords: language learning; linguistic complexity; readability; video difficulty

NOW INTEGRAL TO MODERN LANGUAGE instruction, the use of digital videotexts in language teaching and learning reflects a widespread global trend: On YouTube alone, over one billion hours of material is watched per day across more than 100 countries in 80 different languages (YouTube, n.d.). Language educators have long recognized the role of videotexts in providing authentic language input for a variety of contexts and uses (see Vanderplank, 2010, 2019). 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 can acquire not only 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 their current ability, learning is more likely to halt. An automated system of videotext difficulty can help in selecting videotexts for use in language teaching and learning.

Text difficulty assessment concerns how to accurately assess and predict difficulty in texts as

The Modern Language Journal, 106, 2, (2022)
DOI: 10.1111/modl.12773
0026-7902/22/393–410 $1.50/0

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 and scientific approach to the study of text readability only started in the early 1920s. Beginning in earnest with the work of Lively & 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 and tools for estimating reading difficulty, which proved to be useful in many applications (see Benjamin, 2012, for a comprehensive review).

While the use of videotexts in language learning and teaching has recently received much attention, little research has investigated videotext complexity and difficulty. Though part of a larger research agenda, our attention in the current study is focused solely on the language component of videotexts, as a way to examine how and to what extent linguistic complexity contributes to the perceived difficulty of a videotext for second language (L2) learners. Potentially, if it exists, a strong relationship between linguistic complexity and L2 learners' perception of videotext difficulty would inform L2 listening research, pedagogical materials production, and listening assessment design.

For the purposes of this current study, we first reviewed the literature to set out key areas of linguistic complexity research. We then conducted a study of 322 English-as-a-foreign-language (EFL) learner ratings of videotext complexity. In particular, we leveraged recent advances in natural language processing (NLP) tools for extracting and computing a broad array of acoustic, lexical, syntactic, and discoursal complexity indices and investigated their role in explaining and predicting videotext difficulty.

UNDERSTANDING LINGUISTIC COMPLEXITY

Widely understood as a multilayered construct, linguistic complexity encompasses elements of acoustic, phonological, lexical, semantic, syntactic, and discoursal complexity (Bulté & Housen, 2012). Previous research to develop readability estimates has been largely directed at understating the complexity of written texts. Well-known results of these efforts, for example, include the Flesch reading ease formula (Flesch, 1948) and the Dale–Chall formula (Dale & Chall, 1948). Such work relies on lexical sophistication and sentence complexity as a core basis for the measurement of text difficulty. Despite their widespread usage, however, traditional readability formulas have been criticized for lacking strong construct validity (Crossley et al., 2017; Davison & Kantor, 1982); recent work integrates a series of more sophisticated techniques to achieve increasingly fine-grained levels of text analysis (Benjamin, 2012). Similarly, L2 investigations have been concerned with both written (Crossley et al., 2008; Sung et al., 2015) and spoken texts (Kotani & Yoshimi, 2016; Révész & Brunfaut, 2013; Yoon et al., 2016) but have, to date, ignored research with videotexts. Situated in an EFL context, the present study aims to investigate linguistic complexity features including acoustic, phonological, lexical, syntactic, and discourse features, to determine if these features can provide insights into estimates of videotext difficulty.

Acoustic and Phonological Complexity

Speech, especially spontaneous speech, is typically replete with disfluencies such as pauses, hesitations, and false starts. These disfluencies occur up to six times per hundred words of speech (Tree, 1995); once thought to serve no function, recent psycholinguistic research suggests that disfluencies serve rhetorical and communicative purposes that include orienting listener attention (Collard et al., 2008), providing clues about speaker descriptions (Corley & Hartsuiker, 2003), and leading listeners to expect new information (Arnold et al., 2003). Although high-proficiency learners find frequent pauses distracting (Watanabe et al., 2008), they benefit low-proficiency learners. Studies on the role of silent and filled (e.g., "um," "uh") pauses have demonstrated that frequent pauses aid L2 listening comprehension because they allow extra time for processing (Blau, 1990; Buck, 2001).

Another source of difficulty for language learners is fast speech, in that intelligibility appears to decrease with increased speed. Because a higher speech rate makes processing and decoding spoken words much harder (Chang, 2018), L2 listeners often cite fast speech as a major cause of listening difficulties (Flowerdew & Miller, 1992; Graham, 2006). Listeners' ability to recognize words from speech also depends on their knowledge of the intonation system of the learned language (Chun, 2002). Intonation is the variation of speech pitch, and it serves various functions, including marking prominence, focus, or newsworthiness of a piece of information; differentiating between statements and questions; and signaling

sentences and topic boundaries (Chun, 2002). Further, a higher variation in pitch is correlated with perceptions of speaker liveliness (Hincks, 2005).

Phonotactic probability, or the likelihood of occurrence of a sound sequence, is another property of speech that is believed to help listeners recognize words (Vitevitch & Luce, 2004). In English, for instance, the initial phonotactic sequence /spr–/ is possible, whereas /spm–/ is not (Crystal, 1994). Several empirical studies have confirmed the influence of phonotactic probability on first 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. Research has explored the impact of intonation and phonotactic probability on L2 listening (Bradlow & Pisoni, 1999; Révész & Brunfaut, 2013). In one study, Révész and Brunfaut (2013) investigated the relationship between listening difficulty, pitch variability, and phonotactic probability in 18 listening passages, and found no significant interaction among the variables. The fact that all listening passages in their study were narrated by one speaker, however, as the researchers concluded, decreased the probability that any effects were detected.

Lexical Complexity and Psycholinguistic Properties of Words

Lexical complexity is a multidimensional and multifaceted construct that consists of three components: lexical sophistication, lexical density, and lexical diversity (Lu, 2012). Lexical sophistication, or rareness, refers to the proportion of sophisticated or difficult words in a text (Laufer & Nation, 1995). Though a formal definition of what a difficult word consists of has yet to be agreed upon (Daller et al., 2007), word difficulty is often associated to less frequent words on the assumption that they are less likely to be known and, therefore, less likely to be recognized by language learners (Ellis, 2002; Gries & Ellis, 2015).

Another factor, lexical density, refers to the proportion of content words (CWs) to the total number of words in a text (Ure, 1971). The concept has 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). Because CWs (e.g., nouns, verbs, adjectives, and adverbs) generally carry more information in language than function words (e.g., preposition, pronouns, and conjunctions), the related term lexical density is seen to be a measure of information density or packaging (Johansson, 2009). Processing texts with high lexical density exerts a greater cognitive load on language listeners (Bloomfield et al., 2010). Finally, lexical diversity encompasses the range and variety of vocabulary that is used in a text. Intuitively, texts containing a large diversity of unique words are thought to be more difficult (Bloomfield et al., 2010) and studies have found lexical diversity to be a significant predictor of spoken text difficulty (Révész & Brunfaut, 2013; Rupp et al., 2001).

Cognitive scientists and psycholinguists have long been interested in exploring which characteristics of words affect their processing and learnability (Grainger, 1990; Whaley, 1978). Among several word properties, the psycholinguistic properties of concreteness, meaningfulness, imageability, word familiarity, and age of acquisition have been investigated in the text complexity and readability literature (McNamara et al., 2014). The first of these properties, concreteness, refers to the extent to which content words in a text refer to concrete objects or events as opposed to abstract concepts or ideas (Brysbaert et al., 2014); for example, a word is said to be concrete if one can simply point to the object it signifies (e.g., "chair" or "apple"), and a word is abstract if it can be only described by other words (e.g., "happiness" or "problem"). There is ample psycholinguistic evidence that shows processing abstract words is more challenging than concept words (Paivio et al., 1994).

A second feature, meaningfulness, refers to how closely a particular word is related to other words based on human judgments (Toglia & Battig, 1978); for example, "people" has a high meaningfulness score with human raters, whereas the word "adze" has a very low meaningfulness score in that it is often weakly associated with other words. Further, word imageability refers to how easy it is to create a mental image of the word. The feature is closely related to lexical frequency as they both describe how commonly a word is experienced. A final psycholinguistic property, the age of acquisition, is derived from a score based on human judgments of the age at which a particular word is learned (Kuperman et al., 2012).

Syntactic Complexity

Syntactic processing is an essential component in listening comprehension (Rost, 2011). Therefore, it is reasonable to assume that texts with

greater syntactic complexity are associated with increased listening difficulty (Révész & Brunfaut, 2013). Syntactic complexity refers to the variety and sophistication of syntactic structures in a text, and it has been measured through different metrics in L2 research. Whereas the detrimental effect of syntactic complexity is well documented on written text comprehension, existing work on spoken language difficulty has yielded contradictory findings. For instance, although Révész and Brunfaut (2013) found no impact of syntactic complexity on listening difficulty, Yoon et al. (2016) found that the average number of words per sentence and the frequency of long sentences are good predictors for estimating listening difficulty.

Discourse Complexity

In typical speech, speakers seldom organize complex thoughts into single utterances; instead, they develop their thoughts and expand upon them over many utterances in the discourse (Buck, 2001). Like lexical and syntactic complexity, cohesion is believed to be involved in the determination of text ease or difficulty (McNamara et al., 1996; Sheehan et al., 2010). Cohesion is commonly defined as the explicit characteristics of the text that assist readers or listeners in connecting ideas within the text (Halliday & Matthiessen, 2014). A common cohesive device is the use of connectives (e.g., "therefore," "because"), which tie text sentences and paragraphs together while giving the reader or listener cues about the types of relationships that exist between the different ideas and concepts in the text (Halliday & Matthiessen, 2014). Less cohesive texts are assumed to be more difficult because readers or listeners must invoke more mental resources to fill in the gaps in the text (Kintsch, 1998; Zwaan & Radvansky, 1998). When there are few or no gaps, a text seamlessly moves from one point of information to another, while giving the listeners all the help they need to build new knowledge.

To reiterate, different aspects of language have been identified as being more difficult or challenging for L2 learners. While previous research has primarily focused on written and spoken texts, the present study sets out to investigate linguistic complexity in videotexts. In particular, the study addresses two key research questions:

RQ1. Can linguistic complexity features explain and predict video lecture difficulty?

And if so,

RQ2. Which linguistic complexity features contribute the most to the prediction of video lecture difficulty?

METHOD

Video Corpus

For the purposes of our larger research agenda, we built a corpus of 640 videos in a corpus named *Second Language Videotext Complexity* (SLVC; Alghamdi et al., 2021). In the present study, we made use of only 320 videotexts from this corpus that deploy different instructional designs, as per typologies of video lecture instructional design provided by Crook & Schofield (2017). The selected video lectures were delivered by native speakers of English across discipline areas in the humanities, social studies, education, computer science, mathematics, business and management, and life and medical sciences. Table 1 provides descriptive information about the corpus.

Rating Instrument. For rating the difficulty of the video lectures in our corpus, we adopted and modified a scale that was originally developed for rating L2 listening difficulty (Yoon et al., 2016). Our modified scale comprised five questions (see Appendix). The first question asks for a rating of an overall understanding of a video lecture. The second question asks about short-term retention of the lecture material, and the next two questions concern acoustic and lexical characteristics. The last question asks the participants to rate their overall perception of video difficulty in relation to their language ability.

Though the scale covers different aspects of video lecture difficulty, it does not include questions addressing, for example, grammatical, syntactic, or visual complexity. The scale was designed to be combined into a single composite score during analysis, and we followed the same scoring procedures as in Yoon et al. (2016). Particularly, each questionnaire item is designed to be scored from 1 to 5. Therefore, the highest possible number of points is 25 and the lowest possible score is 5. We deliberately kept the scale short to avoid respondent burden, which may result in low-quality responses (Graf, 2008). Another operational decision we made was to provide the scale in the first language, as opposed to English, to reduce possible misunderstandings due to language problems. Because we are interested in developing a single score for video

TABLE 1
Descriptive Statistics of the Video Lecture Corpus

Discipline	Videos	Duration in Minutes (mean)	Words (mean)
Biology	27	255 (9.47)	35,000 (1,302)
Business	13	52 (4.02)	8,000 (608)
Computer Science	29	165 (5.72)	21,000 (739)
Economics	28	164 (5.72)	24,000 (841)
Education	14	145 (10.38)	27,000 (1,912)
English	74	291 (3.94)	43,000 (584)
History	104	1015 (9.76)	172,000 (1,650)
Mathematics	31	106 (3.42)	16,000 (525)
All	320	2197 (6.58)	346,000 (1,020)

difficulty, responses are averaged for each video. The modified scale was piloted, and the results of a Cronbach's alpha analysis showed a satisfactory internal consistency among the instrument items: $\alpha = .81$, above the suggested benchmark value of .70 (Nunnally, 1978). Post hoc analysis on all participants' responses showed a higher reliability score (items = 2826; $\alpha = .86$).

Raters. For the main study, we asked first-year college students who were enrolled in an intensive English course to participate in a study to rate the difficulty of the videotexts in the SLVC subcorpus. A total of 322 students volunteered to participate and completed the approved human research ethics protocols. To control for the effect of language ability, we ensured that the language learners had taken the Cambridge English Placement Test (CEPT) and placed at the B1 level (Intermediate) of the Common European Framework of Reference (CEFR; Council of Europe, 2001).

Rating Procedure. Upon arrival to a computer lab, participants were randomly assigned to rate a cluster of 10 videotexts from one of the 64 clusters we had created for the rating sessions, with three participants evaluating each cluster. Clusters were designed to be split into two sets of five videotexts each to allow participants time for a break between rating sessions. Ratings were made possible by showing the videotext alongside the rating questionnaire on a single website. Before starting the rating session, participants were informed that they could pause at any time during the session and resume whenever they were ready, to minimize the effect of fatigue on their ratings (Graf, 2008). The majority of the 322 participants completed the two rating sessions, with 10 full videotexts total, over a period of approximately 45 minutes.

Extracting and Computing Linguistic Complexity Features

Acoustic and phonological complexity features were extracted from the audio streams of the videos, whereas lexical, syntactic, and discourse cohesion features were extracted from the phonotactic transcription of the videos. To transcribe our videos, we used Microsoft's speech recognition technology.[1] Upon retrieving the transcribed data, we manually checked all transcripts and corrected mistakes where appropriate. Then, we used the official Stanford NLP Python package (Qi et al., 2019) to parse all video transcripts in our dataset. Taking a running transcript (text) as input, the parser segmented the video transcript into a list of individual sentences and words (tokens) and generated the base forms of all words (lemmas), their parts of speech, morphological features, and syntactic dependencies between words in the sentences. Based on the parser output, we computed features related to lexical frequency and mean dependency distance (MDD). We also used the NLP tool TAALES (Kyle et al., 2018) to compute features related to *n*-gram strength of association, the tool TAACO (Crossley et al., 2019) for measuring discourse cohesion, and Syntactic Complexity Analyzer (Lu, 2010) for computing syntactic complexity.

Phonation Time, Speed, and Pauses. We computed the phonation time as the proportion of time spent talking. The delivery of speech was assessed through three metrics: speech rate, articulation rate, and average syllable duration. 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). Average syllable duration is the phonation time divided by the number of syllables. Pauses are periods of silence exceeding 250 milliseconds (Révész & Brunfaut, 2013). Pitch was calculated using the measure of the fundamental frequency (F0), which estimates vocal cord vibrations per second in voiced sounds. All the features were computed by Praat software (Boersma & Weenink, 2019) using a script developed by de Jong & Wempe (2009).

Lexical Frequency and Sophistication. Lexical frequency was assessed using a band-based approach (Laufer & Nation, 1995; Morris & Cobb, 2004), as it is more appropriate for receptive lexical knowledge and more easily interpretable (Crossley et al., 2013). Specifically, we developed a Python script to calculate the percentages of all words (AWs), function words (FWs), and CWs belonging to the 1,000 (K1), 2,000 (K2), and 3,000 (K3) most frequent word families (groups of semantically related words with the same root, e.g., "nation," "nationalize," "nationalization," etc.) in BNC/COCA lists of word families. We also calculated the percentage of words included in the 2,000 band (K1 + K2 words), and the 3,000 band (K1 + K2 + K3 words).

Whereas FWs are closed-class grammatical words (e.g., prepositions), there is notable variability in how CWs have been defined in previous studies (O'Loughlin, 1995; Ure, 1971). For this study, we adopted the work of Lu (2012), who defined CWs to consist of

> 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') (p. 192)

Lexical sophistication was operationalized as words, nouns, and verbs that are not in the first 2,000 most frequent lemmatized word families.

Formulaic Sequence. In addition to general single word frequency, we also calculated the proportion of formulaic sequences that appeared in the Academic Formulas List (Simpson–Vlach & Ellis, 2010) and Academic Spoken Word List (Dang et al., 2017). The Academic Formulas List comprises 600 multiword units that are pervasive in academic language and the Academic Spoken Word List consists of 1,741 word families that were selected from an academic spoken corpus.

Psycholinguistic Norms. Among the various psycholinguistic attributes of words, our study examined whether concreteness, imageability, and age of acquisition of content words were good predictors of video lecture difficulty. We used TAALES to compute indices related to concreteness, familiarity, meaningfulness, and age of acquisition. TAALES derived these measures from the MRC Psycholinguistic Database (Coltheart, 1981) that contains (a) human ratings of more than 150,000 words on 26 psychological properties, (b) concreteness norms collected for 37,058 lemmas and 2,896 bigrams by Brysbaert et al. (2014), and (c) the age of acquisition norms collected for 30,000 lemmas by Kuperman et al. (2012).

N-*Gram Strength of Association.* Association measures of *n*-grams (i.e., continuous sequences of words) assess the degree to which words in *n*-grams occur together. For example, *n*-grams such as "would be" would have a higher strength of association score than *n*-grams such as "great the." Two well-attested association measures are mutual information and *t*-score (Church & Hanks, 1990; Hunston, 2002). Both measures compare how often *n*-grams appear in a corpus with how often they would be predicted to appear based on the frequency of the words that compose them. Mutual information tends to highlight *n*-grams made up of low-frequency words whereas the *t*-score tends to highlight those made up of high-frequency words (Bestgen & Granger, 2014). Other approaches to word association strength are based on the directionality of the association, in that they quantify whether one word is more predictive of a second word or the other way around (Gries, 2013). A well-established directional measure is ΔP, which calculates the probability of word occurrence based on a cue (another word). The score is calculated using the formula $P(O|C) - P(O|-C)$—that is, ΔP is the probability of an outcome given a cue minus the probability of an outcome without the cue (Kyle et al., 2018). We then used TAALES to compute association strength for both bigrams and trigrams.

Lexical Diversity and Density. Lexical diversity refers to 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 (TTR). The TTR index is based on a simple ratio of types to all tokens in a text. One problem of TTR is that its values are affected by text length, which means that as the number of tokens increases, the likelihood of those tokens being unique decreases. More recently, several approaches to assess lexical diversity have been proposed, each purporting to

measure lexical diversity while having little or no effect on text length. Examples of such approaches are measure of textual lexical diversity (MTLD) and vocd-D, which use estimation algorithms (McCarthy & Jarvis, 2010). However, Covington & McFall (2010) proposed the moving-average type–token ratio (MATTR) as an alternative to TTR that is independent of text length. In a recent study, Treffers–Daller et al. (2018) found that traditional measures of lexical diversity (e.g., TTR and the Index of Guiraud) were more effective in discriminating between texts of different CEFR levels than more sophisticated measures (e.g., D, HD-D, and MTLD), provided text length was kept constant.

As these approaches gauge lexical diversity in different ways, McCarthy & Jarvis (2010) recommended the use of multiple indices instead of only a single index that may not capture the full range of aspects of lexical diversity. In light of this suggestion and the fact that measures have not been explored for video difficulty before, we used both traditional (e.g., root TTR) and more recent measures of lexical diversity (e.g., D, HD-D, and MTLD) for AWs, CWs, and FWs. Lexical density was operationalized as the proportion of all CWs to running words in the video transcripts.

Syntactic Complexity. Using the syntactic complexity analyzer, we computed 14 syntactic complexity indices proposed by Lu (2010). The indices measure the length of syntactic structures (clause, T-unit, and sentence), amount of subordination, amount of coordination, degree of phrasal sophistication, and overall sentence complexity. Lu argued that these indices align with the four dimensions of syntactic complexity suggested by Norris & Ortega (2009) and hence, they can be expected to capture syntactic variety and sophistication (Lu, 2017).

In addition to these features, we computed the MDD (Liu, 2007; Liu et al., 2017) in the video transcripts. 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" (Liu et al., 2017, p. 172). For example, the dependencies between words in a figure can be adjacent or nonadjacent—that is, the two words forming a dependency may appear next to each other or separated by intervening words. Dependency distance is seen to be an important index of memory burden (i.e., dependency locality theory; Gibson, 1998) 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. Liu (2007) proposed two formulas to calculate the MDD of a sentence and a text:

$$\text{MDD (sentence)} = \frac{1}{n-1} \sum_{i=1}^{n} |DD_i|$$

where n is the number of words in the sentence and DD_i is the dependency distance of the *i-th* syntactic link of the sentence, and

$$\text{MDD (the text)} = \frac{1}{n-s} \sum_{i=1}^{n} |DD_i|$$

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

Discourse Cohesion. Among the various discoursal complexity features, we 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.

We used the NLP tool TAACO, version 2.0 (Crossley et al., 2019) to measure local and overall cohesion in video transcripts. In particular, we 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 et al., 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.

TABLE 2
Top Four Correlated Acoustic, Lexical, Syntactic, and Discoursal Features with Participants' Ratings of Video Lecture Difficulty

Category	Feature	Mean	*SD*	*r*
Acoustic	Phonation time	338.29	200.51	.52
	Frequency of pauses	98.71	62.92	.36
	Pitch variation	65.64	14.88	–.34
	Speech rate	3.68	.65	.32
Lexical	BNC/COCA K3 (AW)	7.45	3.76	.64
	BNC/COCA K1 (AW)	68.26	10.32	–.60
	Age of acquisition (CW)	6.27	.61	.56
	Lexical density (type)	.66	.09	.55
Syntactic	Complex nominals (clause)	1.06	.32	.49
	Mean length of clause	9.24	1.86	.46
	Complex nominals (T-unit)	1.97	.91	.39
	Mean length of T-unit	17.07	6.55	.33
Discourse	Coordinating conjuncts	.01	.01	–.51
	Basic connectives	.05	.01	–.43
	All causal connectives	.02	.01	–.43
	All demonstratives	.03	.01	–.38

Note. BNC/COCA = British National Corpus and Corpus of Contemporary American English; AW = all words; CW = content words; K1 = 1,000 most frequent word family; K3 = 3,000 most frequent word family.

Data Analysis

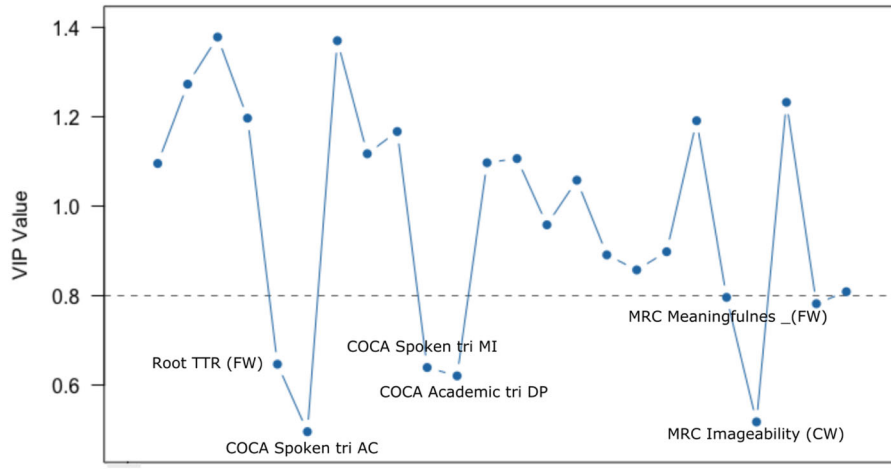
Exploratory Data Analysis. After extracting and computing the linguistic complexity features ($n = 168$), an explanatory data analysis was performed on the dataset. First, missing data were imputed with the feature mean, and features with low variance were removed because they do not contribute much information. The remaining features were then standardized, and correlational analyses between those features and participants' ratings of video lecture difficulty were conducted. The results showed that 130 of the linguistic complexity features significantly correlated with the participants' ratings of video difficulty ($p < .001$) and above $r = .1$ (Cohen, 1988). Table 2 shows the most highly correlated acoustic, lexical, syntactic, and discoursal features with participants' ratings of video lecture difficulty.

Model Training and Evaluation. We used partial least squares regression (PLS-R) to develop a model that predicts video lecture difficulty. PLS-R is a predictive technique, and it is often considered to be a better alternative to ordinary least squares regression, especially when the number of predictors is larger than the sample and there is a risk of multicollinearity (Abdi, 2010; Wold et al., 1984). PLS-R reduces the number of features to a smaller set of uncorrelated components (also known as latent variables or factors) in a manner similar to principal component analysis (PCA). The major difference is that PCA components are determined solely by *x* data, whereas PLS-R components are constructed based on both *x* and *y* (Abdi, 2010). Specifically, the PLS-R algorithm constructs components that maximize the covariance between *x* and *y*.

Determining the optimal number of components in PLS-R is crucial; retaining less than the ideal number of components implies that the model is underfitting the training data and there is still information left that can be modeled, whereas choosing too many components decreases the model interpretability and leads to overfitting (Wiklund et al., 2007). A common approach to determine the number of components in PLS-R is through cross-validation. In this study, the number of components of the PLS-R model was selected based on a randomized test, a technique proposed by Wiklund et al. (2007).

To develop a parsimonious model with minimal overfitting, irrelevant and noisy features should be removed (Mehmood et al., 2012). Variable importance in projection (VIP) is a technique that provides an estimate of the contribution of each predictor to the PLS model (Wold et al., 1993) and is thus commonly used to select predictors. A predictor with a higher value of VIP score is more relevant to predict the dependent variable. Normally, the average of the squared values of VIPs for all predictors is in PLS-R; thus, VIP values greater than 1 indicate predictors that are more important to the model (Eriksson et al., 2013). Because data structures are

FIGURE 1
Variable Importance in Projection (VIP) Plot
[Color figure can be viewed at wileyonlinelibrary.com]



Note. AC = approximate collexeme strength; COCA = Corpus of Contemporary American English; CW = content words; DP = Delta P association score; FW = function words; MI = mutual information; MRC = MRC Psycholinguistic Database (Coltheart, 1981); TTR = type–token ratio.

TABLE 3
Cumulative Explained Variance by Eight Extracted Components and Their Prediction Accuracy

Component	*X* Explained Variance	*Y* Explained Variance	R^2	RMSE
1	54.9	43.78	0.44	2.73
2	68.4	53.22	0.53	2.49
3	74.1	55.24	0.55	2.44
4	78.4	55.67	0.56	2.42
5	82.9	55.79	0.56	2.42
6	87.5	55.86	0.56	2.42
7	91.7	55.89	0.56	2.42
8	93.1	55.96	0.56	2.42

Note. RMSE = root mean squared error.

diverse, the cutoff threshold of 1 may not be optimal for all types of data structures (Mehmood et al., 2012). Therefore, model statistics such as regression coefficients and loading weights are commonly used alongside VIP scores to help select predictors (Mehmood et al., 2012). In the current study, a less conservative VIP score of 0.8 was chosen as a cutoff value (Sawatsky et al., 2015). Figure 1 shows the VIP scores associated with each predictor before removing less important features.

After exploring different solutions, the best PLS-R model has four components that collectively explained 78% of the variance in the *x*-data matrix and 56% of the variance in the *y* vector. It should be noted that the inclusion of the fifth, sixth, seventh, and eighth components increased explained variance in both *x* and *y* spaces but did not increase prediction accuracy significantly (see Table 3), mainly because later components capture more noise than signal in the data. Figure 2 shows the regression coefficients of the pruned model.

To compare the predictive performance of the PLS-R model, the most widely used readability index—namely, the Flesch reading ease formula (Flesch & Gould, 1949)—was used as a baseline model. 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. Different evaluation metrics were utilized to compare the two models: Pearson's correlation coefficient (between the observed and predicted

FIGURE 2
Regression Coefficients and Their 95% Confidence Intervals [Color figure can be viewed at wileyonlinelibrary.com]

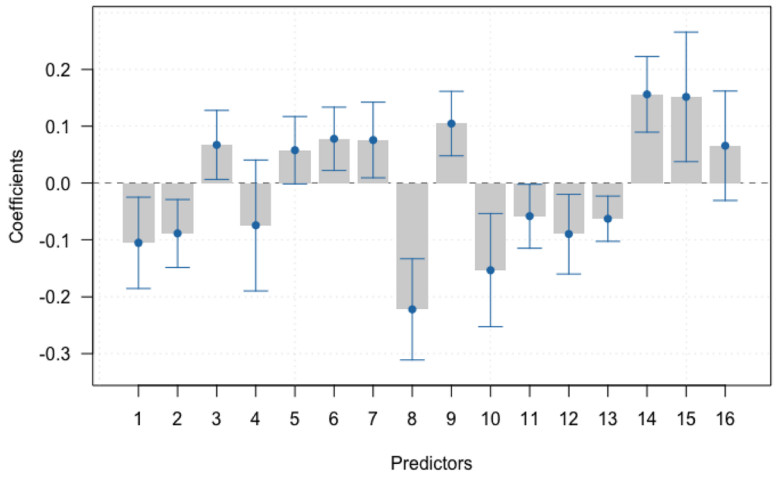


TABLE 4
Comparison of Models' Performance in Predicting Video Lecture Difficulty

Model	R^2	r	RMSE
Baseline model (out of sample)	0.02	.47	3.46
PLS-R (out of sample)	0.52	.72	2.47
Baseline model (in-sample)	–0.93	.19	3.56
PLS-R (in-sample)	0.53	.73	2.50

Note. RMSE = root mean squared error; PLS-R = partial least squares regression.

video difficulty scores) and root mean squared error (RMSE). The RMSE metric is computed using the following formula:

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

Descriptive and correlational analyses were performed in R using RStudio (RStudio Team, 2016), and the multivariate data analysis package in R was used to train and validate the PLS-R model (Kucheryavskiy, 2020).

RESULTS

Regression Models

The prediction performances of the PLS-R model and baseline model were evaluated on a held-out testing set (n = 106) using a ten-fold cross-validation approach. The result demonstrated that the PLS-R model substantially outperformed the baseline model that employed the Flesch reading ease formula in explaining and predicting the difficulty of video lectures in a held-out testing set (see Table 4). The PLS-R model explained 52% of the variance in video difficulty and yielded an RMSE of 2.47 whereas the baseline model explained only 2% of the variance in video difficulty and had an RMSE of 3.46. When it was cross-validated on the entire dataset, the PLS-R model showed a stable prediction performance (R^2 = .53, RMSE = 2.50) compared to the baseline model (R^2 = –.93, RMSE = 3.56).

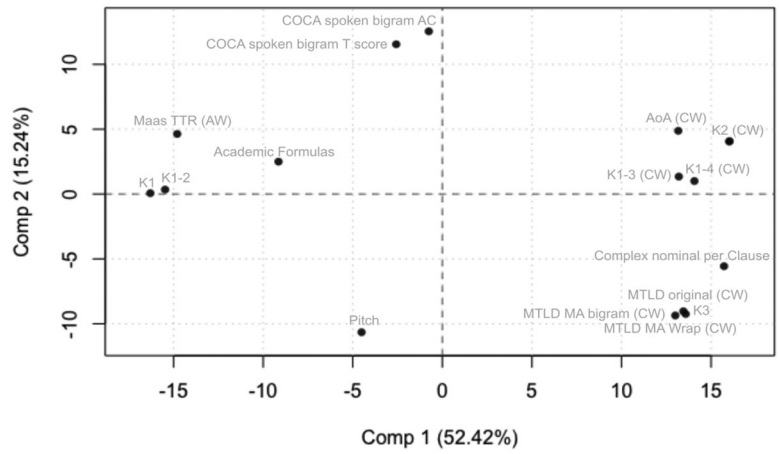
To examine the relative contribution of linguistic complexity to overall prediction, the regression coefficients were examined (see Table 5). The results showed that variability in pitch and frequency from the Academic Formulas List had the highest negative coefficients whereas words' age of acquisition and the index of spoken bigram strength association (t-score) had the highest positive coefficients. The loadings plot for the first two components of this model, shown in Figure 3, suggests that variation in pitch is negatively correlated with video lecture difficulty

TABLE 5
Estimated Regression Coefficients and VIP Scores for All Features in the PLS-R Model

	Feature	Regression Coefficient	VIP
1	Pitch variation	–0.11	0.96
2	BNC/COCA K1	–0.09	1.15
3	BNC/COCA K3	0.07	1.25
4	BNC/COCA K1–K2	–0.07	1.08
5	BNC/COCA K2 (CW)	0.06	1.25
6	BNC/COCA K1–K3 (CW)	0.08	1.00
7	BNC/COCA K1–K4 (CW)	0.08	1.05
8	Academic formulas	–0.22	0.97
9	Complex nominal per clause	0.10	1.01
10	Maas TTR (AW)	–0.15	1.02
11	MTLD original (CW)	–0.06	0.82
12	MTLD MA bigram (CW)	–0.09	0.82
13	MTLD MA Wrap (CW)	–0.06	0.83
14	Kuperman AoA (CW)	0.16	1.14
15	COCA spoken bigram *t*-score	0.15	0.71
16	COCA spoken bigram AC	0.07	0.71

Note. AoA = age of acquisition; AC = approximate collexeme strength; AW = all words; BNC/COCA = British National Corpus and Corpus of Contemporary American English; CW = content words; K1 = 1,000 most frequent word family; K2 = 2,000 most frequent word family; K3 = 3,000 most frequent word family; MA = moving average; MTLD = measure of textual lexical diversity; PLS-R = partial least squares regression; TTR = type–token ratio; VIP = Variable importance in projection.

FIGURE 3
Feature Loadings on the First Versus Second Component



Note. AoA = age of acquisition; AC = approximate collexeme strength; AW = all words; COCA = Corpus of Contemporary American English; CW = content words; MA = moving average; MTLD = measure of textual lexical diversity; K1 = 1,000 most frequent word family; K2 = 2,000 most frequent word family; K3 = 3,000 most frequent word family; TTR = type–token ratio.

(negative value in both components) and that the percentages of BNC/COCA K2, K1–K3, and K1–K4 in video lectures are positively correlated with video difficulty (positive values in both components).

DISCUSSION

The twofold purpose of this study was (a) to investigate whether linguistic complexity can be used to predict video lecture difficulty, and (b) if

so, to examine the relative contributions of acoustical and phonological, lexical, syntactic, and discourse complexity to the prediction of video difficulty. To extend the foundations of previous work, the linguistic complexity in 320 video lectures was analyzed using a contemporary set of computational linguistic tools to extract a comprehensive variety of linguistic complexity features from the audio streams and phonotactic transcripts. The linguistic complexity features were selected based on research from areas of L2 acquisition, psycholinguistics, and discourse comprehension. We now discuss the findings in relation to our research questions.

Explaining and Predicting Video Difficulty Using Linguistic Complexity Features

In our first RQ, we asked whether language learners' perception of video lecture difficulty could be explained and predicted using linguistic complexity features. To this end, we developed a model using PLS-R. Our results demonstrated that a significant proportion of the variance in video lecture difficulty ($R^2 = .52$) was explained by acoustical, lexical, and syntactic features and that there was no significant contribution of textual cohesion. Using a tenfold cross-validation approach, the findings showed that the prediction accuracy of the PLS-R model was consistent and that the model outperformed a baseline model based on the Flesch reading ease index by a significant margin. Finally, the use of traditional measures of text complexity, such as the Flesch reading ease formula, does not provide an accurate estimate of video lecture difficulty.

The Relative Contribution of Acoustic, Lexical, Syntactic, and Discoursal Complexity to Video Difficulty

The second RQ of the current study asked what the relative contributions of acoustic, lexical, syntactic, and discoursal complexity are to the prediction of video lecture difficulty. The results demonstrated that of the 10 features of acoustical and phonological complexity investigated in the study, only pitch variability (pooled $M = 65.64$, $SD = 14.88$) was a significant predictor of video difficulty. Specifically, we found that video difficulty tends to decrease when the speakers have high pitch variation. Previous work has shown that speech with considerable pitch variability is more enjoyable, understandable, and helps listeners to recall information (Hincks, 2005). The remaining indices of acoustical and phonological complexity such as speech rate, articulation rate, pausing, and phonotactic probability had no significant effect on video difficulty. Similar results were reported in previous studies (Révész & Brunfaut, 2013).

As for lexical complexity, our findings accord with our expectation that greater lexical complexity would be associated with increased difficulty. Our analysis showed that lexical frequency, diversity, strength of association, and age of acquisition contributed to the prediction of video difficulty in our dataset. Specifically, we found a relationship between word frequency and video difficulty; that is, if video lectures had a large proportion of less frequent words (K3 and K4), they were rated more difficult. For adequate listening comprehension, language listeners need to know between 2,000 and 3,000 words (Van Zeeland & Schmitt, 2013). We also observed that video lectures that employed many formulaic sequences (e.g., "it turns out that" and "if you look at the") were less difficult for language learners. With regard to psycholinguistic word information, a single psycholinguistic word attribute—age of acquisition—was a significant predictor of video difficulty.

Further, our results concerning the contribution of lexical diversity to video lecture difficulty contradict the general findings in the literature, which suggest there is a significant link between lexical diversity and the difficulty of spoken text (Révész & Brunfaut, 2013; Rupp et al., 2001). Specifically, our results showed that lexical complexity, as determined by three variants of MTLD, negatively correlated with video difficulty. That is, the more lexically diverse the video lecture is, the easier it is perceived by language learners. Considering the fact that approximately 80% of vocabulary in our video lectures was from the first 2,000 most frequent words in the BNC/COCA corpus—which are supposedly known by language learners at the B1 level (Dang et al., 2020)—it is reasonable to believe that our participants benefited from listening to the videos' narrators explaining a concept using different familiar words.

As far as the complexity of the video syntactic structure is concerned, our results showed that a higher ratio of complex nominals per clause in relation to the overall number of clauses in a video text is a key predictor of difficulty. Previous studies have generally observed no significant effect of structural complexity on listening

difficulty (Blau, 1990; Kostin, 2004; Révész & Brunfaut, 2013). The result can be partly attributed to the small-sized datasets used in these studies, which made revealing useful patterns in the text syntax difficult. The fact that the difficulty of video lectures tends to increase in the presence of complex nominals in the video text is highly anticipated. Complex nominals are syntactic constructions that include nouns plus, for example, an adjective, possessive, prepositional phrase, adjective clause, nominal clauses, and gerunds and infinitives in subject position (Cooper, 1976; Lu, 2011). Because complex nominals come before the main verb, they potentially place heavier demands on working memory. Additionally, there is ample evidence suggesting a relationship between the number of words before the main verb—or left-embeddedness—and an increased processing difficulty (Gibson, 1998; Just & Carpenter, 1999)

Last, with regard to the relationship between discourse cohesion and video difficulty, we found that discourse cohesion does not contribute to overall video difficulty prediction. This finding is in line with the results of Nissan et al. (1996) and with those of Révész & Brunfaut (2013), who observed no association between discourse connectives and difficulty in listening to spoken text.

IMPLICATIONS FOR RESEARCH AND PRACTICE

Taken together, the study provides insights about what makes a video lecture difficult for language learners and may contribute to the development of language learning research, materials, and assessment instruments. Recommendations touch on seven points.

Recommendations for Research

1. We have shown that linguistic complexity has a detrimental effect on language learners' perceptions of the difficulty of a video lecture. Therefore, we recommend that researchers take into consideration the impact of linguistic complexity when conducting video-based studies or when developing video tools for language learners.
2. We also found that the Flesch reading ease formula is not a reliable measure of video lecture difficulty. We believe that this finding also applies to other similar traditional measures, for example, the simple measure of gobbledygook (SMOG) and Gunning fog index, which also estimate text complexity based on simple indices of sentence and word difficulty.
3. Our regression model shows a promising result, and we encourage researchers to further explore linguistic complexity and its contribution to video lecture difficulty. Also, our features can be easily computed using existing NLP and speech-processing tools and therefore can be used in developing automated systems for video lecture difficulty assessment and prediction.

Recommendations for Educational Practice

1. Language instructors and video content designers should be aware that L2 learners encounter language challenges when learning from video lectures. To mitigate this effect, we recommend educators use common and less sophisticated words and avoid complex syntactic structures in their video lectures so that video content becomes more accessible for learners.
2. We found that L2 learners experience less difficulty if 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 speak clearly, with proper intonation and articulation.
3. Because 14 of the 16 linguistic complexity features that made the four components in the PLS-R model are related to lexical complexity and sophistication, we recommend that language teachers pay more attention to word difficulty in video lectures. Specifically, teachers should select video lectures that do not have many words that are uncommon.
4. Finally, we found that a greater diversity of words from the K1–K2 BNC/COCA corpus is more helpful for language learners at the B1 level.

LIMITATIONS AND FUTURE WORK

Though our study provides useful insights, limitations must be acknowledged and addressed in future research. First, in this study, we presumed a linear relationship between linguistic complexity and video difficulty; however, the relationship between the two may not be strictly linear and

is likely affected by a number of complicating factors (e.g., topic familiarity). Second, although we created a corpus of academic video lectures, it is by no means fully representative of the genre nor especially large: Millions and millions of videotexts have been made available, and the number increases each day (YouTube, n.d.). Third, our participants were all intermediate EFL students who had the same ethnic and language background; therefore, our findings may not apply to other populations of L2 learners. Finally, though beyond the scope of the present study, we did not address the role of visual complexity (i.e., the visual features of the video, such as the graphics and presence of the instructor on-screen); further, we did not explore the impact of elements of instructional video design, for example, on video difficulty (Fiorella et al., 2019).

As we pursue an overall agenda, we will continue to conduct research in three areas: First, we plan to evaluate the generalizability of our linguistic complexity feature set on other genres of videotext, for example, government advertisements, television shows, and movies. Second, we plan to investigate the relation of visual complexity to instructional video design as a predictor of video lecture difficulty using visual complexity tools such as (AUVANA; Alghamdi et al., 2021). Finally, we will incorporate both linguistic and visual complexity features in an automated web tool for assessing video lecture difficulty and evaluate its utility in pedagogical activities.

CONCLUSION

In this study, we explored the relative contributions of linguistic complexity to the difficulty of video lectures as experienced by English language learners. We evaluated the predictive power of a wide range of linguistic complexity features, and our regression model explained 52% of the variance in video difficulty. When compared to a baseline model, the performance of our model outperformed the baseline model by a significant margin. Overall, the findings of our study provide evidence of the relationship between linguistic complexity and L2 learners' perception of videotext difficulty and that linguistic complexity can be used to explain and predict videotext difficulty. As we begin to better understand what makes a video lecture difficult, we believe that our research provides a promising starting point for future research to develop useful applications akin to those that have already been produced to estimate text readability.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial and technical support of the King Abdulaziz University, Jeddah (DSR grant D-1008-126-1443). Additionally, the authors would like to thank the University of Melbourne for providing support to allow our work to be published as open access.

Open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australian University Librarians.

NOTE

[1] Microsoft's speech-to-text service can be accessed using the following URL: https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/

Open Research Badges

This article has earned Open Data badge. Data this available at https://www.iris-database.org.

REFERENCES

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdisciplinary Reviews: Computational Statistics, 2*, 97–106.

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

Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal there, um, new information. *Journal of Psycholinguistic Research, 32*, 25–36. https://doi.org/10.1023/A:1021980931292

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review, 24*, 63–88. https://doi.org/10.1007/s10648-011-9181-8

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

Blau, E. K. (1990). The effect of syntax, speed, and pauses on listening comprehension. *TESOL*

Quarterly, *24*, 746–743. https://doi.org/10.2307/3587129

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2010). *What makes listening difficult? Factors affecting second language listening comprehension* (Technical Report TTO 81434 E.3.1). University of Maryland Center for Advanced Study of Language. Accessed 28 February 2022 at https://apps.dtic.mil/sti/pdfs/ADA550176.pdf

Boersma, P., & Weenink, D. (2019). Praat: Doing phonetics by computer. Accessed 9 August 2019 at http://www.praat.org/.

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*, 2074–2085. https://doi.org/10.1121/1.427952

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

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity; accuracy and fluency in SLA* (pp. 21–46). Benjamins.

Chang, A. C.–S. (2018). Speech rate second language listening. *The TESOL Encyclopedia of English Language Teaching*, *1968*, 1–7. https://doi.org/10.1002/9781118784235.eelt0580

Chun, D. (2002). *Discourse intonation in L2: From theory and research to practice*. John Benjamins.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*, 22–29.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

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*, 696–702. https://doi.org/10.1037/0278-7393.34.3.696

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*, 497–505. https://doi.org/10.1080/14640748108400805

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

Corley, M., & Hartsuiker, R. J. (2003). Hesitation in speech can… um… help a listener understand. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Behaviour* (pp. 276–281).

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Press Syndicate of the University of Cambridge.

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

Crook, C., & Schofield, L. (2017). The video lecture. *Internet and Higher Education*, *34*, 56–64. https://doi.org/10.1016/j.iheduc.2017.05.003

Crossley, S., 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*, 965–981. https://doi.org/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*, 475–493. https://doi.org/10.1002/j.1545-7249.2008.tb001

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*, 14–27. https://doi.org/10.3758/s13428-018-1142-4

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*, 340–359. https://doi.org/10.1080/0163853X.2017.1296264

Crystal, D. (1994). *An encyclopedic dictionary of language and languages*. Penguin.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin*, *27*, 37–54. https://doi.org/10.2753/JEI0021-3624440403

Daller, H., Milton, J., & Treffers–Daller, J. (2007). *Modelling and assessing vocabulary knowledge*. Cambridge University Press.

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. *Language Learning*, *67*, 959–997. https://doi.org/10.1111/lang.12253

Dang, T. N. Y., Webb, S., & Coxhead, A. (2020). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research*, Advance online publication. https://doi.org/10.1177/2F1362168820911189

Davies, M. (2010). The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*, 447–464. https://doi.org/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*, 187–209.

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*, 385–390. https://doi.org/10.3758/BRM.41.2.385

Ellis, N. (2002). Frequency effects in language processing. A review with implications for theories of im-

plicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*, 143–188. https://doi.org/10.1017/S0272263102002024

Eriksson, L., Byrne, T., Johansson, E., Trygg, J., & Vikström, C. (2013). *Multi-and megavariate data analysis basic principles and applications* (Vol. 1). Umetrics Academy.

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*, *12*, 895–906. https://doi.org/10.1037/edu0000408

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221–233.

Flesch, R., & Gould, A. J. (1949). *The art of readable writing* (Vol. 8). Harper.

Flowerdew, J., & Miller, L. (1992). Student perceptions, problems and strategies in second language lecture comprehension. *RELC Journal*, *23*, 60–80. https://doi.org/10.1177/2F003368829202300205

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*, 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

Graf, I. (2008). Respondent burden. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 739–740). https://doi.org/10.4135/9781412963947

Graham, S. (2006). Listening comprehension: The learners' perspective. *System*, *34*, 165–182. https://doi.org/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*, 228–244. https://doi.org/10.1016/0749-596X(90)90074-A

Gries, S. T. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics*, *18*, 137–166. https://doi.org/10.1075/ijcl.18.1.09gri

Gries, S. T., & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning*, *65*, 228–255. https://doi.org/10.1111/lang.12119

Guichon, N., & McLornan, S. (2008). The effects of multimodality on L2 learners: Implications for CALL resource design. *System*, *36*, 85–93. https://doi.org/10.1016/j.system.2007 .11.005

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's introduction to functional grammar*. Routledge.

Hincks, R. (2005). Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism. *System*, *33*, 575–591. https://doi.org/10.1016/j.system.2005.04.002

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524773

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing. *Lund Working Papers in Linguistics*, *53*, 61–79.

Just, M. A., & Carpenter, P. A. (1999). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122–149.

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

Kostin, I. (2004). Exploring item characteristics that are related to the difficulty of TOEFL dialogue items. *ETS Research Report Series*, *2004*, i-59. https://doi.org/10.1002/j.2333-8504.2004.tb01938.x

Kotani, K., & Yoshimi, T. (2016). Effectiveness of linguistic and learner features to listenability measurement using a decision tree classifier. *The Journal of Information and Systems in Education*, *16*, 7–11. https://doi.org/10.12937/ejsise.16.7

Kucheryavskiy, S. (2020). Mdatools—R package for chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *198*, 103937. https://doi.org/10.1016/j.chemolab.2020.103937

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

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*, 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical density in L2 written production. *Applied Linguistics*, *16*, 307–322. https://doi.org/10.1093/applin/16.3.307

Leech, G., & Rayson, P. (2014). *Word frequencies in written and spoken English: Based on the British national corpus*. Routledge.

Lively, B. A., & Pressey, S. L. (1923). A method for measuring the vocabulary burden of textbooks. *Educational Administration and Supervision*, *9*, 389–398.

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. https://doi.org/10.1016/j.plrev.2017.03.002

Lorge, I. (1944). Predicting readability. *Teachers College Record*, *45*, 404–419.

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

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*, 36–62. https://doi.org/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*, 190–208. https://doi.org/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*, 493–511. https://doi.org/10.1177/0265532217710675

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, *78*, 91–121. https://doi.org/10.1016/S0010-0277(00)00109-8

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

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*, 1–43. https://doi.org/10.1207/s1532690xci1401

McQueen, J. M., & Pitt, M. A. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language*, *39*, 347–370. https://doi.org/10.1006/jmla.1998.2571

Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *118*, 62–69. https://doi.org/10.1016/j.chemolab.2012.07.010

Mintz, T. H., Walker, R. L., Welday, A., & Kidd, C. (2018). Infants' sensitivity to vowel harmony and its role in segmenting speech. *Cognition*, *171*, 95–107. https://doi.org/10.1016/j.cognition.2017.10.020

Morris, L., & Cobb, T. (2004). Vocabulary profiles as predictors of the academic performance of teaching English as a second language trainees. *System*, *32*, 75–87. https://doi.org/10.1016/j.system.2003.05.001

Nissan, S., DeVincenzi, F., & Tang, K. L. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension* (TOEFL Research Report No. RR-51). Educational Testing Service.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*, 555–578. https://doi.org/10.1093/applin/amp044

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw–Hill.

O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing*, *12*, 217–237. https://doi.org/10.1177/026553229501200205

Paivio, A., Walsh, M., & Bons, T. (1994). Concreteness effects on memory: When and why? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1196–1204. https://doi.org/10.1037/0278-7393.20.5.1196

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). Association for Computational Linguistics.

Révész, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, *35*, 31–65. https://doi.org/10.1017/S0272263112000678

Rost, M. (2011). *Teaching and researching listening* (2nd ed.). Longman.

RStudio Team. (2016). *RStudio: Integrated development for R [computer software]*. RStudio, Inc.

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*, 185–216. https://doi.org/10.1080/15305058.2001.9669470

Sanders, T. J. M., Spooren, W. P. M., & Noordman, L. G. M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, *15*, 1–35. https://doi.org/10.1080/01638539209544800

Sawatsky, M. L., Clyde, M., & Meek, F. (2015). Partial least squares regression in the social sciences. *The Quantitative Methods for Psychology*, *11*, 52–62. https://doi.org/10.20982/tqmp.11.2.p052

Sheehan, K. M., Kostin, I., & Futagi, Y. (2010). Generating automated text complexity classifications that are aligned with targeted text complexity standards. *ETS Research Report Series*, *2010*, i–44. https://doi.org/10.1002/j.2333-8504.2010.tb02235.x

Simpson–Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*, 487–512. https://doi.org/10.1093/applin/amp058

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*, 371–391. https://doi.org/10.1111/modl.12213

Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms*. Lawrence Erlbaum.

Tree, J. E. F. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, *34*, 709–738. https://doi.org/10.1006/jmla.1995.1032

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*, 302–327. https://doi.org/10.1093/applin/amw009

Ure, J. (1971). Lexical density: A computational technique and some findings. In M. Coultard (Ed.), *Talking about text* (pp. 27–48). English Language Research, University of Birmingham.

Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension?. *Applied Linguistics*, *34*, 457–479. https://doi.org/10.1093/applin/ams074

Vanderplank, R. (2010). Déjà vu? A decade of research on language laboratories, television and video in language learning. *Language Teaching*, *43*, 1–37. https://doi.org/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*, 407–423. https://doi.org/10.1080/09571736.2019.1610033

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, *36*, 481–487. https://doi.org/10.3758/BF03195594

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*, 81–94. https://doi.org/10.1016/j.specom.2007.06.002

Whaley, C. P. (1978). Word–nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, *17*, 143–154. https://doi.org/10.1016/S0022-5371(78)90110-X

Wiklund, S., Nilsson, D., Eriksson, L., Sjöström, M., Wold, S., & Faber, K. (2007). A randomization test for PLS component selection. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *21*, 427–439.

Wold, S., Johansson, E., & Cocchi, M. (1993). PLS—Partial least squares projections to latent structures. In H. Kubinyi (Ed.), *3D QSAR in drug design, theory, methods, and applications*. ESCOM Science Publishers.

Wold, S., Ruhe, A., Wold, H., & Dunn, W. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, *5*, 735–743.

Yoon, S.-Y., 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). Association for Computational Linguistics.

YouTube. (n.d.). *YouTube for Press*. https://blog.youtube/press/

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*, 162–185. https://doi.org/10.1037/0033-2909.123.2.162

APPENDIX

Second Language Listening Difficulty Scale (modified from Yoon et al., 2016)

How would you rate your understanding of the video?

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

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

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

Estimate the number of words you missed or did not understand.

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

The speech rate was . . .

5 - fast
4 - somewhat fast
3 - neither fast nor slow
2 - somewhat slow
1 - slow

I believe the video is . . .

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

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.