JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 5(3), 182–197. http://dx.doi.org/10.18608/jla.2018.53.12

Impact of Lecturer's Discourse for Student Video Interactions: Video Learning Analytics Case Study of MOOCs

Thushari Atapattu[1], Katrina Falkner[2]

ABSTRACT

Lecture videos are amongst the most widely used instructional methods within present Massive Open Online Courses (MOOCs) and other digital educational platforms. As the main form of instruction, student engagement behaviour, including interaction with videos, directly impacts the student success or failure and accordingly, in-video dropouts positively correlate with dropout from MOOCs. The primary focus of previous video learning analytics studies is on analyzing video interaction behaviour using *explicit* factors (i.e., views). Limited research studies focus on *implicit* video learning analytics (e.g., pause, seek, content type) and their impact on student success, with existing studies addressing video interactions and their relationship with visual transitions. We aim to explore the association between video interactions and non-visual (i.e., verbal) content. This research focuses on language and discourse features of lecture video contents. We conduct a fine-grained analysis of 3.4 million video interactions across two AdelaideX MOOCs — Programming (Code101x) and Cyberwar, Surveillance and Security (Cyber101x). According to our results, a number of discourse features (e.g., lexical diversity and causal connectives) demonstrate a statistically significant correlation with video interactions. We present insights regarding educational video design implications based on discourse processing theories.

Notes for Practice

- Previously, limited research studies focus on the impact of implicit video learning analytics (e.g., pause, seek, contents) on student success within MOOCs. Accordingly, researchers found a relationship between student video interactions and visual transitions of videos.
- However, the impact of verbal contents (i.e., lecturer's verbal discourse) on student video interactions is not yet being explored.
- This research explores the relationship between student video interactions and discourse features of video transcripts.
- Our results demonstrate that discourse features — like first- or second-person pronouns, concrete words, frequent content words, causal connectives, and narrativity — facilitate video discourse processing while lengthy sentences, lengthy videos, and high speaking rate provide difficulties with video discourse processing.
- The findings of this research provide insightful implications for MOOC staff and video producers on formulating video transcripts to reduce interaction peaks.

Keywords

Video analytics, MOOCs, discourse, NLP, Coh-metrix.

Submitted: 13.09.2017 — **Accepted:** 30.05.2018 — **Published:** 11.12.2018

Corresponding author [1]Email: thushari.atapattu@adelaide.edu.au Address: School of Computer Science, University of Adelaide, Adelaide, SA 5005, Australia, ORCID ID: https://orcid.org/0000-0002-0632-4482

[2] Email: katrina.falkner@adelaide.edu.au Address: School of Computer Science, University of Adelaide, Adelaide, SA 5005, Australia, ORCID ID: https://orcid.org/0000-0003-0309-4332

1. INTRODUCTION

Lecture videos are amongst the core pedagogical components of many Massive Open Online Courses (MOOCs; (Breslow et al., 2013; Diwanji, Simon, Märki, Korkut, & Dornberger, 2014; J. Kim et al., 2014; Sinha, Jermann, Li, & Dillenbourg, 2014) and other digital educational platforms, such as Khan Academy, Crash Course, and VSauce (Hua, 2015). During the past decade, Khan Academy has produced 440 million free lecture videos for millions of online subscribers (Hua, 2015). The

massive amount of video learning analytics being generated from digital educational platforms provides a key data source to facilitate the exploration of video engagement patterns of online learners. It has been established that student engagement behaviour with MOOC content (e.g., lecture videos, discussion forums) is likely to predict student success in the course (Kizilcec, Piech, & Schneider, 2013; Wang & Baker, 2015). According to Halawa, Greene, and Mitchell (2014) and Sinha et al. (2014), dropping out from videos (i.e., exiting the video without watching it completely) is positively correlated with dropout from MOOCs.

The primary focus of previous video learning analytics studies is on how student performance is affected by video engagement behaviour using *explicit* factors such as views and annotations (de Barba, Kennedy, & Ainley, 2016; Hsin & Cigas, 2013; Risko, Foulsham, Dawson, & Kingstone, 2013). Related studies utilize the two terms *engagement* and *video interaction* interchangeably to refer to activities related to clickstreams, in-video quizzes, views, etc. The analysis conducted in this paper refers to "video interactions" as implicit video analytics (e.g., clickstreams such as pause, seek, play, etc.). Recent research (Guo, Kim, & Rubin, 2014; Zhang, Zhou, Briggs, & Nunamaker, 2006) found that students are more likely to engage when the videos: 1) are brief (i.e., shorter than 6 minutes); 2) include PowerPoint slides and talking heads; 3) include Khan-style tablet drawing; 4) are recorded in an informal setting (e.g., office) instead of a studio; and 5) embed interactive activities like in-video quizzes. However, in-video dropout remains a challenging issue irrespective of the adoption of these video design considerations. Further, MOOC participants are likely to exponentially decline in their video engagement as they progress throughout the course (Halawa, Greene et al., 2014). These issues motivate the growing need to explore *implicit* video learning analytics such as clickstream data (e.g., when pausing videos, duration of re-watching the video, changing the speed) and the effect of video contents (i.e., visual and verbal aspects) on the above-mentioned problematic engagement behaviour.

Limited research studies have focused on *implicit* video learning analytics (Giannakos, Chorianopoulos, & Chrisochoides, 2015; Guo, Kim et al., 2014; J. Kim, Li, Cai, Gajos, & Miller, 2014; Li, Kidzinski, Jermann, & Dillenbourg, 2015; Sinha, Jermann et al., 2014). Li et al. (2015) explored the association between video interaction patterns and the perceived video difficulty and found that patterns such as infrequent/large skips, infrequent/large amount of re-watching, and pause frequency indicate higher video difficulty. Sinha et al. (2014) predict video and course dropout using the student's sequence of video clicks. Kim, Guo et al. (2014) explore the rationale behind a significant number of students demonstrating similar video interaction patterns (i.e., peaks) and how this behaviour can be explained by visual contents. Their extensive study found that 61% of peaks are associated with visual transitions (e.g., changing from whiteboard explanation to a talking head, often corresponding to the beginning of new material).

However, the results found by Kim, Guo et al. (2014) cannot be adopted within MOOC lecture videos where less visual transition or visual content is included. For instance, videos in the dataset of Cyber101x rarely contain visual transitions. Additionally, Kim, Li et al. (2014) claimed that 39% of interaction peaks are caused by non-visual transitions (i.e., verbal). Our work is building on this existing research to explore the association between video interactions and lecturers' video discourse. We construct our hypothesis thusly: "*Lecturers' video discourse will correlate with interaction patterns of MOOC videos*." We collectively analyze 3.4 million video interaction records from two courses: Introductory Programming (Code101x) and Cyberwar, Surveillance and Security (Cyber101x), both offered by AdelaideX.[1] Our text corpus contains 128 video transcript files, with a total of 4,710 sentences.

2. BACKGROUND

2.1. Video Learning Analytics

Video learning analytics (or educational video analytics) is an emerging subfield of learning analytics research, focusing on collecting, analyzing, and reporting data about learner interaction with videos, which allows educators, researchers, and instructional designers to better understand and improve video-based learning and teaching (Mirriahi & Vigentini, 2017). Video learning analytics gained interest among the learning analytics (LA), educational data mining (EDM), and human–computer interaction (HCI) communities over the last decade with the increasing popularity of digital educational platforms (e.g., Khan Academy) and MOOCs. The applications of video learning analytics can be categorized into two strands: 1) video content and 2) video usage (Mirriahi & Vigentini, 2017). *Video content analysis* techniques include visual pattern recognition, video indexing, navigation (e.g., annotating important segments; (Risko, Foulsham et al., 2013)), and summarization (Grigoras, Charvillat, & Douze, 2002). *Video usage* analytics can be further classified as *explicit* and *implicit* analytics. The majority of early works focus on *explicit* factors such as views, votes, and their impact on the learning outcome (de Barba, Kennedy et al., 2016; Halawa, Greene et al., 2014; Hsin & Cigas, 2013; Risko, Foulsham et al., 2013; Sinha, Jermann et al., 2014; Wang & Baker, 2015). Additionally, tools like edX insight[2] and YouTube analytics[3] provide dashboards, presenting explicit video data such as views/votes over time.

[1] https://www.edx.org/school/adelaidex
[2] https://insights.edx.org/
[3] https://www.youtube.com/analytics

Although these explicit video analytics provide implications on video usage and popularity, they lack in-depth video interaction information including when the learner is engaged, distracted, or confused. Limited research studies focus on *implicit* video learning analytics such as pause, seek, and stop events (Giannakos, Chorianopoulos et al., 2015; Guo, Kim et al., 2014; J. Kim, Li et al., 2014; Li, Kidzinski et al., 2015; Sinha, Jermann et al., 2014).

Giannakos et al. (2015) developed an open access video learning analytics system to collect clickstream data of video interactions. The authors investigated the relationship between video navigation (e.g., replay, skip) and learning performance. They found the peaks of video activities (e.g., replay) correspond to answers of assessments. Li et al. (2015) investigated the association between video interaction patterns and perceived video difficulty. Their work embeds an in-video survey at the end of each lecture video in their two MOOCs to collect feedback about the content difficulty. The results demonstrate that patterns such as infrequent/large skips, infrequent/large amount of re-watching, and pause frequency indicates higher video difficulty. Based on video clicking behaviour, Sinha and colleagues (2014) constructed a quantitative Information Processing Index (IPI), which consists of cognitively plausible behaviours, with the aim of providing meaningful interventions for students in real time. The authors first categorize a sequence of students' raw clicks (e.g., play-pause-seek_forward-pause-seek_backward) into higher-level behavioural actions such as *slow watching* (e.g., playing and pausing video content) and *clear concept* (e.g., a combination of seeking backward, indicating struggle). Further, based on IPI, they predict whether the student is retained to the end of the video and course and found that student dropout in the MOOC is 37% less likely if they have one standard deviation higher IPI than average.

Kim, Guo et al. (2014) investigated the rationale behind video *interaction peaks.* An interaction peak is a spike that occurs due to a significant number of students interacting with a common segment of a video. The authors measure the effect of visual contents for interaction peaks. Their visual content analysis utilized an image similarity measure, which calculates the pixel difference between two adjacent frames. They found that 61% of interaction peaks occurred due to *visual transitions* such as the beginning of new material, returning to missed content, and following tutorial steps. Based on the dataset from four edX MOOCs, the authors claim that 39% of interaction peaks are related to non-visual explanation. Guo et al. (2014) conducted an empirical study to investigate how video production (e.g., recorded in the studio or informal setting like a classroom) affects student engagement and found that shorter videos (less than 6 minutes) are more engaging. They also found videos are more effective when they are accompanied by talking heads rather than only PowerPoint slides, when videos are recorded in an informal setting rather than studios, and when Khan-style tablet drawing is adopted (Guo, Kim et al., 2014).

2.2. Discourse Analysis

2.2.1. Theoretical background

According to McNamara and colleagues (2014), the term "discourse" is used as an analysis of language, texts, communication, and social interactions through various communication channels. Computational approaches for *discourse analysis* (known as computational discourse science) is considered an emerging area of interdisciplinary research between computational linguistics, computer science, cognitive psychology, and education, and is designed to study the complex processes and patterns associated with the use of language. Our work adopts a multilevel theoretical framework for discourse processing (Dowell, Graesser, & Cai, 2016; A. C. Graesser, McNamara, & Kulikowich, 2011). The work of Graesser et al. (2011) identifies six levels: words, syntax, explicit textbase, situation model, discourse genre and rhetorical structure, and pragmatic communication. Our work relates to the first five levels of the theoretical framework.

The words or lexicon level incorporates psycholinguistic and lexical databases (e.g., MRC, WordNet; (Coltheart, 1981; Miller, 1995) to obtain psycholinguistic properties (e.g., word concreteness, word frequency, synonyms) of words to determine the difficulty or easiness of discourse processing. For instance, *concrete words* are said to be easier to process and understand than *abstract words*. Similarly, text difficulty increases when less frequently used words (i.e., rare words) are present. Word frequency is obtained from lexical databases such as CELEX, which analyzes the relative frequency of words in public documents per million words. It uses a corpus from the Dutch Centre for Lexical Information (Baayen, Piepenbrock, & Gulikers, 1995) that includes 17.9 million words. The *syntax* of sentences also impacts the difficulty (or ease) of discourse processing. Simple syntactic structures with apparent subject-verb-object, short phrases, and active voice are easier to process, while complex, lengthy syntactic structures with many embedded subordinate clauses are difficult to process.

In contrast to words and syntax levels, *textbase* contains explicit ideas in the text that preserve the *meaning*. The basic units of meaning in the textbase are called *propositions* (van Dijk & Kintsch, 1983). A proposition contains a predicate (e.g., main verb, adjective, connectives) and one or more arguments (e.g., nouns, pronouns, embedded propositions). *Cohesion* relation (or referential cohesion) is an important theoretical construct that measures the overlap in words between units in the textbase, such as propositions, clauses, sentences, and paragraphs (McNamara, Graesser et al., 2014). It provides linguistic clues (e.g., noun overlap, stem overlap, and content word overlap) to make connections between an adjacent pair of sentences. In general, cohesively connected texts are supportive for discourse processing (Kintsch, Kozminsky, Streby, McKoon, & Keenan, 1975; McNamara, Louwerse, McCarthy, & Graesser, 2010). Conversely, cohesion gap (i.e., no overlap of content words in surrounding text) at a textbase level contributes to difficulty in discourse processing. The *situation model*

ISSN 1929-7750 (online).

(or mental model) relates to the deeper level of meaning that goes beyond words (Kintsch, 1998). For example, the situation model may include characters, events, actions, thoughts, and emotion of characters in the narrative text. Situation model mainly corresponds to dimensions such as causation, intentionality, and time (Zwaan & Radvansky, 1998). When one or more of these dimensions of situation model discontinue, a break in cohesion occurs. Accordingly, it is important to have *connectives* (e.g., *because, so, in order to, later*) to facilitate cohesion and discourse processing.

The *discourse genre* and *rhetorical structure* level refer to the category of text such as narration, exposition, persuasion, or description (Pentimonti, Zucker, Justice, & Kaderavek, 2010). According to previous research (A.C. Graesser & Ottati, 1996; Haberlandt & Graesser, 1985), narrative texts are read approximately twice as quickly, and remembered twice as well, as informational texts. Thus, we map these levels to our analysis of individual sentences, adjacent sentences, and full-text analysis to identify the impact of the theoretically grounded language and discourse aspects of teachers on student discourse processing and hence, their interaction with MOOC videos.

2.2.2. Coh-Metrix

Coh-Metrix is held to be the broadest and most sophisticated tool grounded with theories of text and discourse comprehension to analyze the text on multiple levels (i.e., from the word/syntax levels to discourse genre and rhetorical structure; (Dowell, Graesser et al., 2016; McNamara, Graesser et al., 2014). Although the initial focus of Coh-Metrix was on text cohesion and coherence, the current publicly available tool (Coh-Metrix 3.0[4]) provides 108 theoretically grounded and validated linguistic and discourse variables of texts. The tool provides numerical scores for the variables (e.g., incidence scores, ratios). Coh-Metrix analyzes text as either adjacent sentences or all sentences together. Analysis of adjacent sentences is mainly utilized to calculate referential cohesion.

The Coh-Metrix version utilized in this paper categorizes the discourse variables (known as *measures*) as descriptive, text easability, referential cohesion, latent semantic analysis (LSA), lexical diversity, situation model, syntactic complexity, word information, and readability. Coh-Metrix provides relatively simple features of text using descriptive measures such as the number of sentences, the mean number of letters in words. Additionally, it provides complex language and discourse features, such as text easability measures (e.g., narrativity, word concreteness, and deep cohesion), which calculate scores about the difficulty (or easiness) of text for discourse processing. *Deep cohesion* reflects the degree to which the text contains causal (e.g., *because, so*) and intentional connectives (e.g., *in order that, so that*), facilitating the learner to form a more coherent, deeper understanding of the text (A. C. Graesser, McNamara et al., 2011). Referential cohesion measures include calculating the overlap in content words between adjacent sentences (local) or all sentences (global), enabling a linguistic clue that aids the creation of connections between text. Conversely, LSA measures the semantic overlap between adjacent sentences; it indicates how a pair of sentences is conceptually similar to each other. Lexical diversity refers to the unique words in text compared to the total number of words. The situation model measures the level of mental representation that goes beyond the explicit words of a text through the variables such as causal content and intentional content. Further information about the variables used in this study is listed in Table 2 and Section 4.

Coh-metrix has been widely utilized for the evaluation of text quality and difficulty (or ease) in student essays, political documents, and textbooks (more information about applications is included in the next section).

2.2.3. Automated discourse analysis in educational contexts

The automated analysis of discourse within the educational context primarily explores the impact of the language and discourse of students on learning and social interactions. Studies related to "textual discourse and learning" include analysis of student writing, such as essays and discussion posts, and its association with learning performance. Writing-Pal (McNamara et al., 2012), an intelligent tutoring system that provides feedback during essay writing, uses the cohesiveness of an essay as an indicator of writing quality.

The massive amount of content generated within MOOCs provides a rich source of data about student discourse (e.g., discussion contents; (Dowell, Graesser et al., 2016) and teacher discourse (e.g., lecture recordings). Crossley et al. (2015) found that certain language and discourse features of discussion posts (e.g., lexical sophistication, n-grams, cohesion, lengthy, and frequent posts) significantly contribute to the completion of MOOCs. Dowell et al. (2015) explore the academic performance and social position of MOOC participants through their use of language and discourse in discussion forums. The results of their studies show that learners perform significantly better when their language includes a more expository style of discourse, abstract language, and simple syntactic structures.

Thus far, *learners'* language, discourse, and communication have been the primary focus of MOOC-related discourse research in order to understand learners' cognitive, motivational, and social processes (Rosé & Ferschke, 2016), including the authors' previous works (Atapattu & Falkner, 2016; Atapattu, Falkner, & Tarmazdi, 2016). Surprisingly, the effect of *teacher* language and discourse on student video interaction is overlooked. Some studies explore the quality of learning materials and their connection with aspects like performance. McNamara, Kintsch, Songer, and Kintsch (1996) and Varner et al. (2013) explore the impact of the coherence of learning materials on student comprehension of the subject domain. Vega,

[4] http://cohmetrix.com/

Feng, Lehman, Graesser, and D'Mello (2013) explore the influence of text complexity of material on cognitive disengagement during reading exercises.

A preliminary study by Kim, Li et al. (2014) explores the effect of contents such as visual, text, and speech for video interaction peaks. They utilize topic modelling techniques over video transcripts to extract the latent topics of each video and measure the topic transition likelihood over time. The results show that 10% of interaction peaks are associated with a topic transition. However, there is no evidence of a change of instructor's pitch with interaction peaks. A study by Guo et al. (2014) see Section 2.1 for more information) found that students are more engaged with shorter videos when the teacher is speaking fast (e.g., words per minute > 160). However, for longer videos (6–12 minutes), slower-speaking rates are more engaging. The authors' explanations of the results include the fact that fast-speaking teachers convey more energy/enthusiasm, and that the lecture is more likely to be accompanied by a PowerPoint presentation, which might contribute to better student engagement. In contrast, teachers who speak slowly in longer videos might be using whiteboard explanations.

To further explore the automated discourse analysis at scale, interested readers can refer to "discourse analysis" studies in the Learning Analytics and Knowledge (LAK), Educational Data Mining (EDM), Computer-Supported Collaborative Learning (CSCL), and Learning at Scale (LAS) communities.

3. METHODOLOGY

3.1. Data

We analyze the AdelaideX Introductory Programming (Code101x) and Cyberwar, Surveillance and Security (Cyber101x) MOOCs offered in 2015 by the Computer Science School and the Law School at the University of Adelaide respectively. During the initial offering of Code101x, 26,129 participants were registered (active — 13,930; completed more than 50% — 2,137; received a verified certificate — 831). Code101x covers introductory programming concepts (e.g., sequencing, iteration, and selection) in the context of creating artwork and animations with ProcessingJS. The course duration is six weeks, with an average of eight videos per week. The average length of a video is 3.63 minutes, adhering to the design recommendations by Guo et al. (2014) i.e., video length under 6 minutes). Code101x was taught by three lecturers, each sharing approximately one-third of the syllabus. The three lecturers share a similar presentation style (e.g., talking head, background animations, and programming screens).

During the first offering of Cyber101x, 24,200 participants were registered (active — 12,360; completed more than 50% — 3,958; received a verified certificate — 1,425). Cyber101x covers topics on the internet, international law and surveillance, and cybersecurity. The course span of Cyber101x is six weeks, with an average of 13 videos per week. The average length of a video is 4.12 minutes (< 6 min). The course was primarily taught by three lecturers. However, academic/industry experts were also featured each week to provide further knowledge to the participants. The majority of the videos in Cyber101x are narrative-style without *visual transitions*.

We extracted 1.5 and 1.9 million de-identifiable records of video interaction events (e.g., play/pause, load/stop, seek video, speed change, and show/hide closed captions or transcripts)[5] from Code101x and Cyber101x respectively. Our text corpus for discourse analysis contains 128 video transcript files (SubRip Text) with a total of 4,710 sentences and associated timespans.

3.2. Method

Our study aims to answer the following research question: Does the lecturers' video discourse correlate with interaction patterns of MOOC videos? If so, what specific discourse features are they?

3.2.1. Video data processing

The cleaning of our video dataset mainly focuses on removing records without a video identification (e.g., URL) or a *current time* (i.e., the moment of video interaction occurred is not being logged). The remaining records are categorized according to their *video id*. We eliminate the first and last few seconds (5–10 seconds) of each video as they reflect auto-load or stop within the mobile app or web browsers[6] (J. Kim, Guo et al., 2014). In each video, we construct an individual student's (de-identifiable) interaction behaviour using the approach[7] discussed by Kim, Guo et al. (2014). This process enables us to eliminate noisy data such as anonymous user records. From this, we eliminate records when students do not return to the video after a week, assuming that the particular *pause* is not related to curiosity or confusion but suggests that the learner will not return to complete the video. Finally, we analyze *play, pause, seek* (e.g., current time), *speed change,* and *show closed captions/transcripts* events, which could potentially be associated with issues of discourse processing.

Table 1 shows the percentage of each event in our filtered dataset. We exclude load and stop events from the analysis since the majority of these occur during the start and end of the video, possibly as automatic events. The majority of

[5] http://edx.readthedocs.io/projects/devdata/en/latest/internal_data_formats/tracking_logs.html
[6] http://www.youtube.com/yt/playbook/yt-analytics.html#details
[7] https://github.com/edx/insights

remaining interactions correspond to "play-pause" and "seek" interactions, with a small remainder made up of a number of additional events, including speed change and closed caption interactions. Our study focuses on three categories of interactions.

- **Combined-events**: Due to the lack of data availability (Table 1), it is challenging to analyze a number of the possible events, such as *speed change* and *show/hide closed captions*, individually. Therefore, we first consider the category of "combined-events" (i.e., all possible events including *play, pause, seek, speed change*, and *show/hide closed captions*) interactions collectively due to the richness of data to discover interesting patterns. For instance, a sample video considered in Figure 1 demonstrated approximately 4000 interactions at the 1.05 minute mark, which increased the curiosity of course staff to review the visual/verbal contents of the video segment.
- **Pause**: This event provides useful insights about when and why students pause the video and whether this kind of behaviour corresponds to confusion or curiosity.
- **Seek:** This event indicates when students seek backward or forward within the video and may reflect knowledge gaps, or identification of pre-existing knowledge.

Table 1. Events and the Corresponding Percentage of Video Dataset (n = 1,440,468)

Event	Play	Pause	Seek	Speed change	Show/hide closed captions (CC)	Load	Stop
Percentage (%)	43	15	10	2	2	20	8
Analysis	Combined-events (play, pause, seek, speed change, show/hide cc)					Not included	Not included

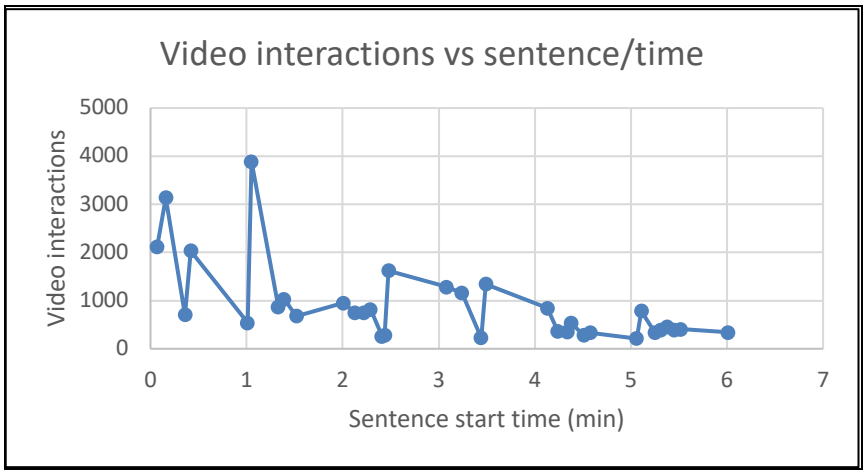

Figure 1. "Combined-events" video interactions against sentence start time in a sample video of Cyber101x.

Unlike the study by Kim, Guo et al. (2014), which considers re-watching behaviour, we only consider the first-time video watching logs of individuals, assuming that re-watching will focus typically on a particular segment of the video, corresponding to further clarifications or linked with activities (e.g., quizzes). We also assume that first-time watchers tend to demonstrate more natural video interactions, which provide a rich source of data about the whole video rather than a particular segment. After constructing an individual student's first-time video interaction timelines, we have integrated all these individual records to examine the video interactions of the whole course (Note: the focus of this study is not on individual student behaviour). Hence, a significant number of students demonstrating similar video interactions (called *interaction peak*; (J. Kim, Guo et al., 2014) provides us with some indications about confusion or interest in the video content.

3.2.2. Discourse processing

In order to apply language and discourse analysis techniques, we extracted sentences from video transcripts. SubRip text (SRT) files contain text and its associated timespan. Since these transcripts are well-written and grammatically complete, we apply NLP techniques directly without any pre-processing. We extracted individual sentences as well as adjacent sentences to compute features like referential cohesion (i.e., overlap of text across sentences).

As discussed in Section 2.2.2, our work utilizes Coh-Metrix 3.0 (McNamara, Graesser et al., 2014) for discourse analysis. Among the 108 theoretically grounded and validated features implemented in Coh-Metrix, we have utilized only the most important features based on our text corpus. For instance, video transcripts are a sequence of sentences and therefore, paragraph-related features are unusable. We were required to re-implement some of the features or reuse algorithms available under GNU General Public License since the tool does not provide an API for sentence-by-sentence analysis and the manual use of the web tool[8] is time-consuming. However, we utilized the Coh-Metrix web tool during complete text analysis (Section 4.3).

Table 2 lists the features used for our analysis.[9] According to McNamara et al. (2014), "measures" are used as theoretical constructs corresponding to discourse processing. We use the term "feature" to assess the "measure." A detailed discussion of these features and their association with video interaction is included in Section 4.

Table 2. Language and Discourse Features Used in our Analysis

Measure	Feature
Descriptive	Word count/sentence length; Syllable count
Text easability principal component scores	Narrativity; Syntactic simplicity; Word concreteness; Referential cohesion; Connectivity
Referential cohesion	Content word overlap
Latent semantic analysis	LSA overlap
Lexical diversity	Type-token ratio; Measure of textual lexical diversity
Connectives	Causal; Logical; Contrastive
Syntactic complexity	Left embeddedness; Number of modifiers per noun phrase
Word information	Pronoun; Frequency for content words
Readability	Flesch Reading Ease
Speaking rate	Sentence length-duration ratio

4. RESULTS

Video interactions could vary based on visual and verbal content. Previous studies by Kim and colleagues (Guo, Kim et al., 2014; J. Kim, Guo et al., 2014) recommended video design implications corresponding to visual transitions. In contrast, this study explores how verbal content (i.e., lecturers' video discourse) contributes to video interaction peaks. This section focuses on reporting the results of our language analysis of verbal content against video interactions. Section 4.1 presents the association between the video interactions and discourse features of each sentence.

4.1. Sentence-by-sentence Analysis

The sentence-by-sentence analysis calculates scores corresponding to the discourse features of each sentence. A simple example is the calculation of word count per sentence. We extract the quantity of video interactions occurring during the timespan of a sentence. Subsequently, we measure the correlation between discourse features and video interactions of all sentences in each course using the Pearson Correlation Coefficient (r) (Table 3).

After obtaining the discourse features that demonstrated a significant correlation with video interactions, we conduct a post-hoc test (i.e., Benjamini-Hochberg procedure; (Benjamini & Hochberg, 1995) to control the false discovery rate (25%). From this analysis, we eliminate variables such as modifiers per noun and pronouns as their significance might be due to chance. The other language and discourse features are retained in the analysis (Table 3).

Descriptive features (DESC) like word count, syllable count, and sentence length demonstrate a positive correlation with video interactions in both courses for "combined-events" as well as specific events like *pause* and *seek*. Figure 2 demonstrates the association between video interactions and word count using a sample video that includes the highest video interactions in Code101x. Please note that the actual video interactions are divided by 50 (known as *frequency*) to align the graph with the sentence word count.

The findings suggest that the longer the sentence, the more likely the students are to interact with the video (e.g., pause or seek), which may be an indicator of their difficulty in processing. For instance, the average word count of sentences in Cyber101x is between 20 and 30. However, one sentence on the topic of an "amendment to legislation" contained 76 words, which indicated an interaction peak. Longer sentences tend to contain complex syntax, which will be discussed further in this section.

[8] http://tool.cohmetrix.com/
[9] http://141.225.42.86/CohMetrixHome/documentation_indices.html

Table 3. Correlation between Discourse Features and Video Interactions

Feature (Category)	Combined-events		Pause		Seek	
	Code	Cyber	Code	Cyber	Code	Cyber
Sentence word count (DESC)	0.388**	0.440**	0.358*	0.496**	0.368**	0.465**
Sentence syllable count (DESC)	0.187**	0.094	0.139*	0.171**	0.151*	0.168**
Sentence length (DESC)	0.370**	0.315**	0.349**	0.381**	0.387**	0.384**
Syntactic simplicity (TexES)	–0.157*	–0.153*	–0.106	–0.132*	–0.174**	–0.164**
Connectivity (TexES)	–0.314**	–0.034	–0.214**	–0.088	–0.290**	–0.021
Causal connectives (Con)	–0.360**	–0.289**	–0.099	–0.320**	–0.267**	–0.382**
Type-token ratio — content words (LD)	–0.303**	–0.203**	–0.274**	–0.269**	–0.319**	–0.218**
Measure of Textual Lexical Diversity (LD)	0.175**	0.162*	0.157*	0.179**	0.145*	0.132*
Left embeddedness (SynCX)	0.146*	–0.010	0.189*	0.019	0.145*	0.011
Frequency for content word (WRD)	–0.184**	–0.013	–0.134*	–0.140*	–0.066	–0.080
Speaking rate (SpD)	0.362**	0.378**	0.294**	0.376**	0.391**	0.138*

***p<0.001, *p<0.05*

Note: Code = Code101x; Cyber = Cyber101x; DESC = Descriptive; TextES = Text easability principal component scores; LD = Lexical diversity; Con = Connectives; SynCX = Syntactic complexity; WRD = Word information; SpD = Speech discourse.

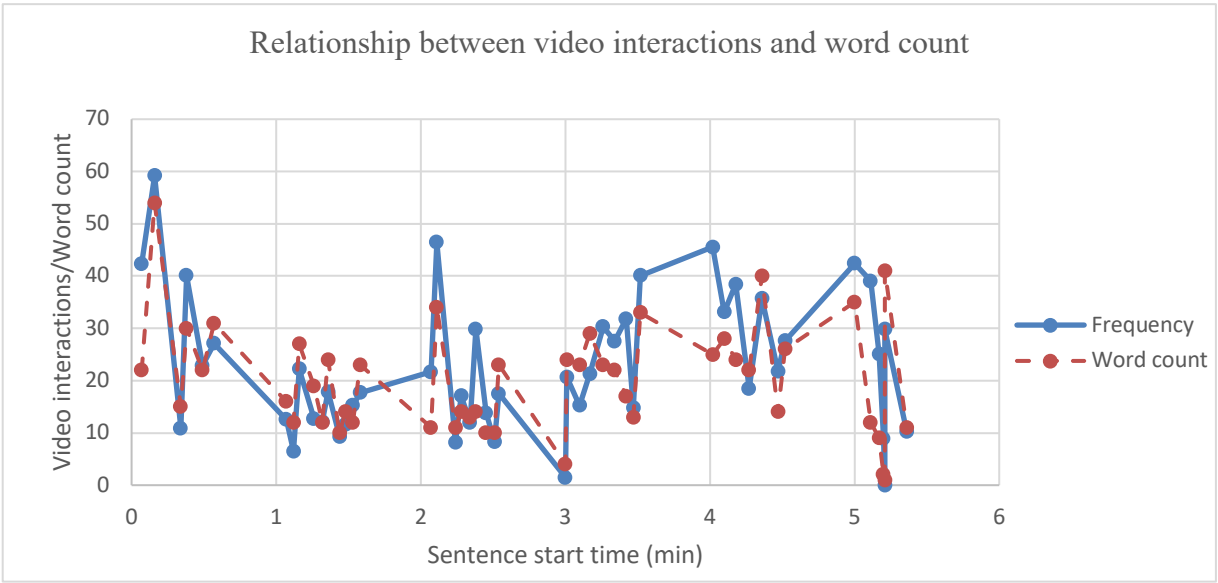

Figure 2. Correlation between "combined-events" video interactions and word count of a sample video in Code101x (r = 0.777); video interactions = frequency * 50.

Text easability features (i.e., *syntactic simplicity, connectivity*) provide insights into the ease (and difficulty) of the text based on its linguistic characteristics. The text easability features are supportive for discourse processing (A. C. Graesser,

McNamara et al., 2011). Accordingly, these features demonstrate a negative correlation with video interactions (Table 3). Our results demonstrate a negative correlation between syntactically simple sentences and video interactions in both courses. For instance, a lesson on "*order of instructions in algorithms*" uses simple and brief sentences like "*now, to draw the circle*," "*you might wonder why it is 255*," "*let's say 220*" to draw shapes using the ProcessingJS. These sentences are less challenging to process, and hence, the learner interaction (e.g., pause) with these particular video segments is relatively low.

Connectivity reflects the cohesive links between ideas and clauses in the sentence using connectives such as logical (*and, or*), additive (*moreover*), contrastive (*however, although*), and causal (*because, so*) (Cain & Nash, 2011; McNamara, Graesser et al., 2014). Even though the connectives create lengthy sentences, the connectivity score relates to faster reading times, better memory, facilitated inference making, and a deeper understanding of the relations in the text (Millis, Golding, & Barker, 1995). According to Table 3, the connectivity incidence score negatively correlates with video interaction, with a significant correlation in Code101x. Our dataset suggests that lengthy sentences (i.e., larger than the average sentence size of the video) *with appropriate connectives* demonstrate fewer video interactions. Conversely, lengthy sentences *without appropriate connectives* tend to demonstrate interaction peaks. This result confirms previous findings on the influence of connectives for processing and comprehension of text.

The results of the study by Cain and Nash (2011) suggest that young readers read text more quickly when two-clause sentences are linked by an appropriate connective (e.g., *causal*) compared to texts in which a connective is either neutral (e.g., *and*), inappropriate to the meaning conveyed by the two-clause, or not present. To explore this, we analyzed common connective categories (e.g., logical, causal, and contrastive) in the discourse (Louwerse, 2001). An interesting finding is observed on *causal* connectives, demonstrating a negative correlation between almost all the event types in both courses (Table 3). Figure 3 demonstrates the association between "seek" video interactions and causal connectives using a sample video that includes the highest seek interactions in Cyber101x. Causal relationships present a consequence, allowing learners to understand the "cause–effect" relation. Our results suggest that inclusion of causal connectives in the academic discourse might improve student discourse processing. Interestingly, some related previous findings on causal connectives suggest that reading times among young readers are quickest when two-clause sentences contain causal connectives rather than adversative and temporal connectives (Cain & Nash, 2011). In contrast, other connectives such as logical and contrastive do not strongly correlate with video interactions. Logical connectives especially, such as "*and*," are very common in sentences (known as *neutral*), producing lengthy sentences that could cause the correlation to be positive but not significant.

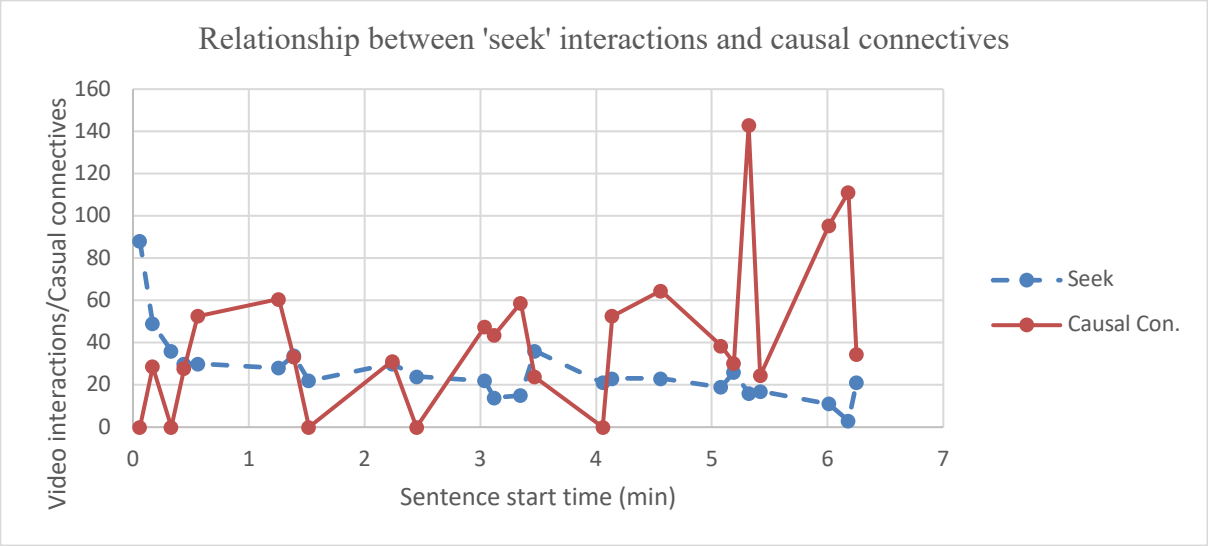

Figure 3. Correlation between causal connectives and "seek" video interactions of a sample video in Cyber101x (r = –0.498).

We obtained a contradictory result for "*lexical diversity*" (LD) features. The most popular lexical diversity calculation is *Type-Token Ratio* (TTR; (Templin, 1957), which measures the ratio between unique words (*types*) and the total number of words (*tokens*) in a sentence. When TTR equals 1, each word occurs only once in the sentence. This increases the difficulty in processing since the sentence is very low in cohesion and the learner has to understand many unique words. For instance, a sentence in Cyber101x with TTR = 0.95 includes many unique "content" words — "*A platform born from the capacity of the internet to facilitate large scale and anonymous transfers of data, it has enabled whistleblowing on a scale unimaginable and practically impossible in the days of photocopiers and typewriters*." When the TTR decreases, words are repeated in the text, making it easier to process and comprehend. Since lexically diverse texts are difficult to process, we expect to obtain a positive correlation with video interactions. Conversely, TTR in Table 3 demonstrates a negative correlation. Our statistical

results demonstrate that lexical diversity negatively correlates with the sentence length. This consequence is also discussed in the work of McCarthy and Jarvis (2010) as they state "*when the number of word tokens increases, there is a lower likelihood of those words being unique."* To overcome this issue, an estimation algorithm is proposed to measure lexical diversity (Measure of Textual Lexical Diversity [MTLD](McCarthy & Jarvis, 2010) , which provides a positive correlation as expected.

Syntactically complex text usually includes many embedded syntax structures in contrast to syntactically simple text with apparent subject-verb-objects (SVO; (Jurafsky & Martin, 2008). According to Graesser, Cai, Louwerse, and Daniel (2006), complex, embedded syntax places heavier demands on working memory. In particular, this is the case when the reader holds many words in the working memory before they receive the main verb of the clause. We measure the syntactic complexity using the features like left embeddedness (i.e., number of words before the main verb) and modifiers per noun phrase. Our results show a positive correlation between syntactic complexity features and video interactions. This suggests that the interaction peaks correlate with the syntactically complex sentence. For instance, a sample sentence ("*In a carefully organized sequence of disclosures through journalist Glenn Greenwald and documentary maker Laura Poitras, Snowden revealed that the …*") has left embeddedness of 17 (i.e., 17 words before the main verb) with video interactions being more than twice the average.

Our analysis also includes the impact of individual words (WRD) in sentences regarding video interaction. In particular, we analyze *pronouns* and *frequent words*. We eliminate "pronouns" in reporting as the post-hoc test indicates the significance might be due to chance. We found that a significant number of learners in the two MOOCs demonstrate low video interactions when the text corpus include "frequent" content words over rare words (word frequency counts of CELEX; (Baayen, Piepenbrock et al., 1995). As discussed in Coh-Metrix (McNamara, Graesser et al., 2014), text difficulty is expected to increase when there are rare words that most learners never or rarely encounter.

We analyzed the *speaking rate* of each sentence in the transcripts, measured as the ratio between sentence length and the time used to deliver that sentence (Note: video speed is set to "1"). We obtained a significantly positive correlation with video interactions. This suggests that high speaking rate often corresponds to more interactions with the video (e.g., *pause-backward_seek-play*), indicating difficulties with discourse processing. According to previous findings (Guo, Kim et al., 2014), the high speaking rate is less problematic if the same content is presented visually in PowerPoint slides. Also, Guo et al. (2014) suggested that speaking relatively quickly with *more energy* and *enthusiasm* is suitable for online videos. Their work suggests that speaking rate is a surface feature that correlates with enthusiasm. Accordingly, an explanation for the positive correlation between video interactions and speaking rate could be due to a lack of visual presentation (using slides) in Cyber101x. Most MOOC platforms (e.g., Edx, Coursera) and video platforms (e.g., YouTube) provide an option for "speed change." Therefore, if the video includes high speaking rate, learners could vary the speed for better comprehension.

Thus, our results demonstrate that certain discourse features are correlated with video interactions. Among them, descriptive features like sentence length and word count demonstrate a positive correlation with video interaction. However, it is likely that longer sentences leave more time for learners to interact, which could impact this high correlation. Therefore, we repeated our analysis by controlling the sentence length and word count. We normalize the video interactions by dividing from word count (and sentence length). The results demonstrated that the significant correlation persisted unchanged for the features other than syntactic simplicity, syntactic complexity, and lexical diversity. However, the features associated with sentence length (i.e., syntactic simplicity or complexity, lexical diversity) do not demonstrate significant correlations when sentence length is controlled. Table 4 demonstrates the correlations for normalized video interactions by controlling the word count. Nevertheless, our analysis considers sentence length (and word count) as features that could impact the video interaction patterns of the learners.

Table 4. Correlation between Discourse Features and Normalized Video Interactions

Feature	Combined-events		Pause		Seek	
	Code	Cyber	Code	Cyber	Code	Cyber
Syntactic simplicity	0.056	0.018	0.108	0.104	0.020	0.022
Connectivity	–0.367**	–0.126*	–0.212**	–0.161*	–0.190**	–0.009
Causal connectives	–0.343**	–0.224**	–0.027	–0.242**	–0.157*	–0.365**
Type-token ratio — content words	–0.086	–0.033	–0.075	–0.070	–0.101	–0.007
Measure of textual lexical diversity	0.011	0.026	0.017	0.015	–0.019	–0.036
Left embeddedness	–0.031	–0.099	–0.044	–0.076	–0.005	–0.091
Frequency for content word	–0.240**	–0.074	–0.179**	–0.127*	–0.076	–0.120
Speaking rate	0.162*	0.349**	0.204**	0.347**	0.269**	0.128*

**p<0.001, *p<0.05

4.2. Adjacent Sentence Analysis

We conducted a separate study to understand the local cohesion of text through the analysis of consecutive, adjacent sentences (i.e., pairs of sentences). This study utilizes discourse features including referential cohesion (local), semantic overlap, and sentence syntax similarity. Our results for adjacent sentence analysis are not significant. We observe some significant correlation in individual videos, but they are not generalizable. For instance, lecturers tend to create cohesive connections in adjacent sentences when they introduce new concepts or topics. A new concept (*Five Eyes Intelligence agreement*) in Cyber101x is repeated in adjacent sentences; however, since this is an unfamiliar concept to the learner when it's first introduced, the sentences that include this concept demonstrate an interaction peak. Additionally, LSA overlap produced insignificant correlations, primarily due to a lack of conceptually similar terms in technical courses like programming and cybersecurity.

4.3. Complete Text Analysis

Text easability features like *narrativity, deep cohesion,* and *referential cohesion (global)* are incoherent when the analysis is performed sentence-by-sentence. Therefore, we calculate the *easability* measures for the whole text in each video transcript. However, video interactions are influenced by the declining nature of video engagement when the course progresses. Therefore, we normalize the video interactions by the number of "active" learners in each video. As previously mentioned, we eliminate the records when learners do not return to the video after one week. The remaining learners (i.e., "active") are used to normalize the video interactions. Table 5 demonstrates the correlation between discourse features and video interactions of active students.

Table 5. Correlation between Discourse Features and Video Interactions of *Active* Students

Feature (Category)	Combined-events		Pause		Seek	
	Code	Cyber	Code	Cyber	Code	Cyber
Sentence count (DESC)	0.372	0.745**	0.446*	0.794**	0.330	0.785**
Word count (DESC)	0.343	0.685**	0.437*	0.735**	0.336	0.719**
Narrativity (TextES)	–0.295	–0.393*	–0.200	–0.420*	–0.204	–0.534**
Word concreteness (TextES)	0.061	0.369	0.114	0.357	0.032	0.412*
Pronouns (WRD)	–0.089	–0.503**	–0.073	–0.543**	–0.099	–0.610**
First-person plural pronouns (WRD)	–0.160	–0.382	–0.022	–0.417*	–0.115	–0.417*
Flesch Reading Ease (RD)	–0.191	–0.543**	–0.028	–0.577**	–0.104	–0.672**

**p<0.001, *p<0.05

Note: Code = Code101x; Cyber = Cyber101x; DESC = Descriptive; TextES = Text easability principal component scores; WRD = Word information; RD = Readability.

As in the study in Section 4.1, we conducted a post-hoc test (i.e., Benjamini-Hochberg procedure) to control the false discovery rate. Accordingly, we eliminate referential cohesion (i.e., *content word overlap*) as its significance might be due to chance. The other language and discourse features retained in the analysis are shown in Table 5.

Descriptive features like sentence count and word count positively correlate with the video interactions. In particular, Cyber101x demonstrated positive a correlation ($p < 0.001$) when combined-events, as well as individual pauses and seek events, are considered. Additionally, Code101x demonstrated a significant correlation between pause events and video interactions. This suggests that the longer the content of the video (i.e., text), the more likely it is that learners perform frequent pauses and seeks in the video. The average number of sentences and word count per video is approximately 37 and 600 in both courses. However, our analysis showed that videos with a high number of sentences (e.g., 83 in Cyber and 58 in Code) demonstrated the most pauses and seeks.

As discussed in Section 4.1, text easability features support discourse processing (A. C. Graesser, McNamara et al., 2011). Therefore, video interactions are expected to correlate negatively with text easability. For instance, highly narrative transcripts/videos are easy to comprehend due to their story-like nature. The results in Table 5 show a negative correlation between narrativity and video interactions in both courses. Word concreteness is another text easability feature that measures whether the text contains more concrete, meaningful words that form mental images (e.g., things, events, properties; (Coltheart, 1981; A. C. Graesser, McNamara et al., 2011; Pennebaker, Booth, & Francis, 2007; Turney, Neuman, Assaf, & Cohen, 2011). If the text contains more concrete words, it would be easier to process and hence would not demonstrate many interaction peaks. Conversely, our results show a significant positive correlation in Cyber101x. This may suggest that higher video interactions relate to abstract text (e.g., ideas and concepts). However, since the results in Code101x are not significant, the impact of *word concreteness* for interaction peaks is not generalizable.

A high density of *pronouns* is ambiguous and can create referential cohesion problems if the learner cannot relate the pronoun to the context (McNamara, Graesser et al., 2014). This suggests that a high use of pronouns can increase video interactions among learners. However, our results are contradictory to this observation. Cyber101x illustrates a significant negative correlation ($p < 0.001$) between pronoun incidence (i.e., the number of words classified as pronouns for a span of 1,000 words) and video interactions (Figure 4). Therefore, we analyzed the dataset further using each category of pronouns (e.g., first-person singular, second person). Both courses contain a low number of first-person singular pronouns (e.g., I). Code101x used first-person singular pronouns to introduce an action conducted using ProcessingJS (e.g., *I'll draw a red circle for the fill*) while in Cyber101x, external experts used "I" to introduce themselves or provide their opinions on cybersecurity concerns. In contrast, both the courses utilized a relatively high number of first-person plural pronouns (e.g., we, us; mean per video is 18 in Cyber101x and 49 in Code101x). The use of "we" and "us" suggests social interaction (Pennebaker, Booth et al., 2007), which helps the learner sense that they are part of a class when engaging with the video (e.g., *We can use variables to help us calculate…*).

Cyber101x demonstrates a significant negative correlation between first-person plural pronouns and video interactions, resulting in fewer pauses and seek events. The second-person pronouns (e.g., you) occasionally occur across video discourse. Also, third-person singular (e.g., he, she) and plural (e.g., they, those) were very rare in both courses (mean per video is less than 10). Third-person pronouns can be more ambiguous to resolve than first- and second-person pronouns. First- and second-person pronouns are straightforward to relate to the context and facilitate discourse processing. Thus, our contradictory outcome of a negative correlation between pronouns and video interactions could occur due to high use of first- and second-person pronouns in the video discourse.

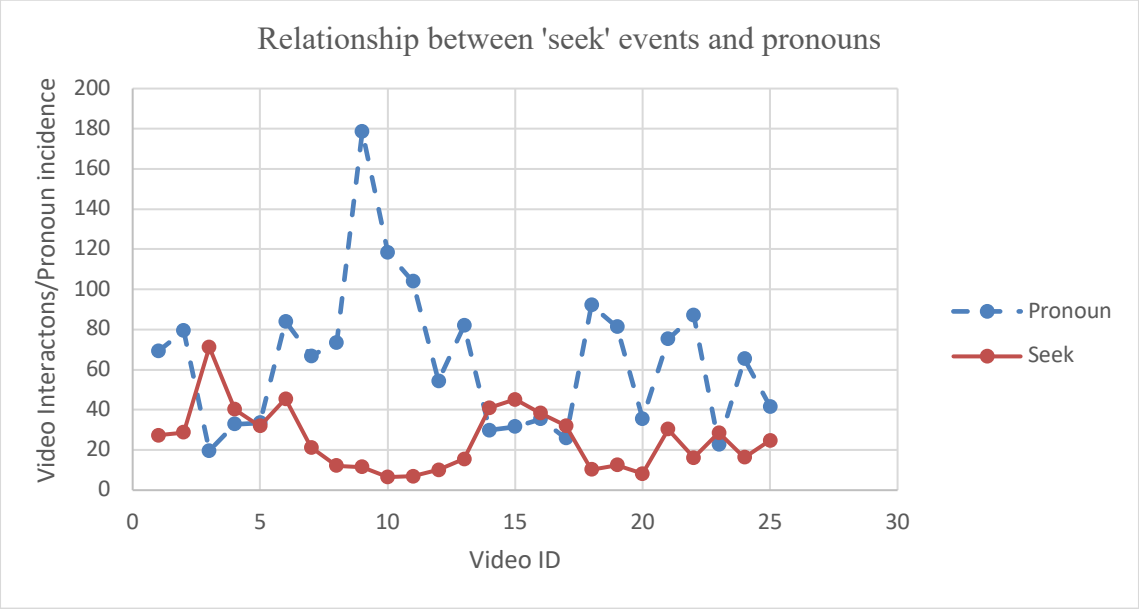

Figure 4. Correlation between pronouns and "seek" video interactions of first 25 videos of Cyber101x; seek interactions are normalized by dividing by active students.

Table 6. Correlation between Discourse Features and Normalized Video Interactions of *Active* Students

Feature	Combined-events		Pause		Seek	
	Code	Cyber	Code	Cyber	Code	Cyber
Narrativity	–0.143	–0.417*	–0.327	–0.433*	–0.316	–0.500**
Word concreteness	0.167	0.414*	0.180	0.431*	0.177	0.158
Pronouns	–0.166	–0.422*	–0.174	–0.484*	–0.063	–0.542**
First-person plural pronouns	–0.236	–0.244	–0.215	–0.472*	–0.176	–0.420*
Flesch Reading Ease	0.059	–0.412*	–0.193	–0.459*	–0.112	–0.484*

***p<0.001, *p<0.05*

Finally, we measure the *readability* of video transcript to scale text on easability or difficulty of video discourse using a popular metric called "Flesch Reading Ease" (Klare, 1974–1975). Flesch Reading Ease is calculated using sentence length and word length. According to Table 5, we obtained a negative correlation between text readability and video interactions for both courses with statistically significant results ($p < 0.001$) in Cyber101x. This result suggests that the higher the readability of text, the more likely it is that learners understand the text, hence video interaction is lower (i.e., negative correlation).

Similar to the sentence-by-sentence analysis (Section 4.1), we normalized the video interactions by controlling for *word count* and repeating the experiments. However, we did not observe any major differences in the patterns; the new figures are reported in Table 6.

4.4. Discussion

Our results demonstrate that descriptive features like sentence count, word count, sentence length, syllable count, and speaking rate positively correlate with video interaction peaks. This suggests that longer texts lead to higher video interactions (e.g., pauses, seeks) since they are more difficult for discourse processing. Our results on sentence and word count are aligned with the suggestion proposed by Guo et al. (2014) on *video length*. Furthermore, according to the results, speaking quickly in lecture video will produce interaction peaks that may occur due to confusion or curiosity. However, as mentioned in Section 4.1, the speaking rate can be adjusted in present video platforms. Our data-driven outcome also supports the importance of syntactically simple, short sentences with apparent subject-verb-object for discourse processing over complex syntactic structures that include many words before the main verb (i.e., left embeddedness). Our results also support the importance of decreasing the lexical diversity of text by reducing the unique words in sentences. Our dataset, particularly Cyber101x, emphasizes the importance of concrete words over abstract words.

In contrast, features that are theoretically supportive for discourse processing (e.g., narrativity, causal connectives) demonstrate negative correlation with video interaction peaks. In general, the increase of "connectivity" in sentences produces longer text. However, our findings suggest that *meaningful connections* such as causal (e.g., *because, so*), intentional (e.g., *in order that, so that*), and contrastive (e.g., *however, but*) connectives are more important in academic discourse, in contrast to logical connectives (e.g., *and, or*) since they scaffold the learner to form more coherent and deeper understanding of the text. Also, ambiguous pronouns can create referential cohesion problems if they cannot be easily related to the context. Therefore, a high density of pronouns in the discourse is predicted as positively correlated with video interactions. However, our results are contradictory to this intuition. As discussed in Section 4.3, this primarily occurred due to the excessive use of first- and second-person pronouns, which are less ambiguous.

Our results also confirmed the importance of using "frequent" content words over "rare" words. According to the results, we also found that highly *readable* sentences demonstrate low video interaction among MOOC participants. Finally, we observed the correlation between *narrativity* and video interactions is significant in Cyber101x, which used more conversational language by introducing people, events, and locations. The preference for conversational language over formal language is supportive for learning, particularly as it fosters a social connection between the teacher and learner (Mayer, 2009). Even though the identified effect of discourse features for video interactions is consistent across the two MOOCs, we also observed some domain-specific interest over some videos in Cyber101x. For instance, Cyber101x delivered some interesting and controversial topics, such as Edward Snowden and Julian Assange. Students tended to highly interact with these videos, irrespective of the language and discourse.

Since our results are consistent across MOOCs from two different disciplines (i.e., technology vs. law) that deliver different kind of content (e.g., worked examples in programming vs content-rich debate and discussion), and also across events such as *pause* and *seek*, we suggest that the patterns revealed through the correlation between video interaction events and certain discourse features provides insights to effectively adapt the video discourse of lecturers.

We acknowledge that other factors can also increase the occurrence of interaction peaks, such as *visual transition*, practising programming using the steps shown in the video in Code101x, and content difficulty or unfamiliarity. Our research has some identified limitations, which could be addressed in future work. Firstly, the current work cannot capture the interactions outside of the context. For instance, we are unable to distinguish whether a *pause* is related to switching between hands-on programming tasks, confusion, or curiosity about the video content. Also, due to the de-identifiable nature of our dataset, the goals of individual learners are unclear. For instance, some students participate in the course to learn only a specific part of the course while others expect to earn a certificate by completing all the course activities. Therefore, we cannot distinguish whether an in-video dropout corresponds to a learner whose desire to learn only a particular segment, thus *skipping* the video. One probable solution would be to collect multi-modal analytics, such as eye tracking or gaze-based data, to identify whether the interactions correspond to disengagement.

Another limitation of our approach is that we do not differentiate between backward or forward seek behaviour. These may be interpreted as differences in intent, e.g., backward seeks may correspond to a knowledge gap or confusion while forward seeks may suggest that the learner possesses the required knowledge. However, the data available is inadequate to support separate analysis of these two behaviours.

Additionally, our work focuses only on *first-time video watchers* in contrast to *re-watching* behaviour as in work by Kim, Guo et al. (2014). We assume first-time video watchers demonstrate natural video engagement behaviour, which provides a

rich source of data about the whole video rather than a particular segment. According to Kim, Li et al. (2014), an interaction peak could occur due to verbal content or visual transition. However, our current analysis did not isolate the interaction peaks caused solely by verbal content by removing the interaction peaks correspond to visual transitions. This differentiation is possible using the techniques developed by Kim, Guo et al. (2014), such as calculating the pixel difference between the two windows.

As the first comprehensive study to analyze the impact of lecturers' video discourse on interaction patterns with MOOC videos, our research provides insightful implications for MOOC course staff and video producers. Based on the outcome of this research, we are able to feed this knowledge back into the MOOC video design, particularly the creation of transcripts and future courses, thus facilitating in an effective discourse comprehension of videos. Additionally, it would be interesting to study whether the data-driven implications provided through this work generalize to face-to-face classrooms.

5. Conclusions and Implications

Our research establishes a step towards understanding hidden patterns of language and discourse that could impact the video interactions of MOOC learners. We analyzed 3.4 million video data items across two MOOCs to understand when a significant number of learners demonstrate similar video interaction patterns (known as *interaction peaks*). Our results show consistent correlations between features of lecturers' video discourse and video interactions. This consistency is observed across events (i.e., *combined* events except for *load/stop* and individual events like *pause* and *seek*) as well as across MOOCs from different disciplines. This research guides the academic staff to pay attention to video interaction peaks that could likely occur due to difficulties with video discourse processing. Further, as a guide for video designers, we summarize the discourse features that can create video interaction peaks:

- Reduce lengthy sentences unless they utilize meaningful causal, contrastive, or intentional connectives.
- Reduce the number of sentences in a video (a.k.a. video length).
- Slow down the speaking rate.
- Use first-person/second-person pronouns over third-person pronouns.
- Use concrete words over abstract words.
- Use frequent content words.
- Use narrative, conversational language over formal language.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors declared no financial support for the research, authorship, and/or publication of this article.

Acknowledgements

The authors would like to thank Katy McDevitt and Tim Cavanagh from the AdelaideX team for providing the data for conducting this research. The authors would also like to thank the MOOC course staff from Cyber101X and Code101X for providing access to their course resources.

References

Atapattu, T., & Falkner, K. (2016). *A Framework for Topic Generation and Labeling from MOOC Discussions*. Paper presented at the Proceedings of the Third ACM Conference on Learning @ Scale, Edinburgh, Scotland, UK. http://dx.doi.org/10.1145/2876034.2893414

Atapattu, T., Falkner, K., & Tarmazdi, H. (2016). *Topic-wise classification of MOOC discussions: A visual analytics approach.* Paper presented at the Proceedings of the 9th International conference on Educational Data Mining, Raleigh, NC, USA.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database*. Paper presented at the Linguistic Data Consortium, Philadelphia, PA.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, 57*(1), 289-300.

Breslow, L. B., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment, 8*, 13-25.

Cain, K., & Nash, H. M. (2011). The influence of connectives on young readers' processing and comprehension of text. *Journal of Educational Psychology, 103* (2), 429 - 441. http://dx.doi.org/10.1037/a0022824

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology, 33*(4), 497-

505. http://dx.doi.org/10.1080/14640748108400805

Crossley, S. A., McNamara, D. S., Baker, R., Wang, Y., Paquette, L., Barnes, T., & Bergner, Y. (2015). *Language to completion: Success in an educational data mining massive open online class.* Paper presented at the Proceedings of the 8th International Conference on Educational Data Mining.

de Barba, P. G., Kennedy, G. E., & Ainley, M. D. (2016). The role of students' motivation and participation in predicting performance in a MOOC. *Journal of Computer Assisted Learning, 32*(3), 218-231. http://dx.doi.org/10.1111/jcal.12130

Diwanji, P., Simon, B. P., Märki, M., Korkut, S., & Dornberger, R. (2014, 13-14 Nov. 2014). *Success factors of online learning videos.* Paper presented at the 2014 International Conference on Interactive Mobile Communication Technologies and Learning

Dowell, N. M., Graesser, A. C., & Cai, Z. (2016). Language and Discourse Analysis with Coh-Metrix: Applications from Educational Material to Learning Environments at Scale. *Journal of Learning Analytics, 3*(3), 72-95. https://dx.doi.org/10.18608/jla.2016.33.5

Dowell, N. M., Skrypnyk, O., Joksimovic, S., Graesser, A. C., Dawson, S., Gaševic, D., . . . Kovanovic, V. (2015). *Modeling Learners' Social Centrality and Performance through Language and Discourse.* Paper presented at the International Conference on Educational Data Mining.

Giannakos, M. N., Chorianopoulos, K., & Chrisochoides, N. (2015). Making sense of video analytics: Lessons learned from clickstream interactions, attitudes, and learning outcome in a video-assisted course. *2015, 16*(1). http://dx.doi/org/10.19173/irrodl.v16i1.1976

Graesser, A. C., Cai, Z., Louwerse, M., & Daniel, F. (2006). Question Understanding Aid (QUAID): A web facility that helps survey methodologists improve the comprehensibility of questions. *Public Opinion Quarterly, 70*, 3–22. https://dx.doi.org/10.1093/poq/nfj012

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing Multilevel Analyses of Text Characteristics *Educational Researcher, 40*(5), 223-234 https://dx.doi.org/10.3102/0013189X11413260

Graesser, A. C., & Ottati, V. (1996). Why stories? Some evidence, questions, and challenges. In R. S. Wyer (Ed.), *Knowledge and memory: The real story* (pp. 121-132). Hillsdale, NJ: Erlbaum.

Grigoras, R., Charvillat, V., & Douze, M. (2002). *Optimizing hypervideo navigation using a Markov decision process approach*. Paper presented at the Proceedings of the tenth ACM international conference on Multimedia, Juan-les-Pins, France. http://dx.doi.org/10.1145/641007.641014

Guo, P. J., Kim, J., & Rubin, R. (2014). *How video production affects student engagement: an empirical study of MOOC videos*. Paper presented at the Proceedings of the first ACM conference on Learning @ scale conference, Atlanta, Georgia, USA. http://dx.doi.org/10.1145/2556325.2566239

Haberlandt, K. F., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology, 114*, 357-374. http://dx.doi.org/10.1037/0096-3445.114.3.357

Halawa, S., Greene, D., & Mitchell, J. (2014). Dropout prediction in MOOCs using learner activity features. *eLearning papers*.

Hsin, W., & Cigas, J. (2013). Short videos improve student learning in online education. *Journal of Computing Sciences in Colleges, 28*(5), 253-259.

Hua, K. (2015). Education as entertainment: YouTube sensations teaching the future., 2016, from http://www.forbes.com/sites/karenhua/2015/06/23/education-as-entertainment-youtube-sensations-teaching-the-future/#57bd75574ca1

Jurafsky, D., & Martin, J. (2008). *Speech and language processing*. Englewood, NJ: Prentice Hall.

Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, K. Z., & Miller, R. C. (2014). *Understanding in-video dropouts and interaction peaks in online lecture videos*. Paper presented at the Proceedings of the first ACM conference on Learning at scale conference, Atlanta, Georgia, USA. http://dx.doi.org/10.1145/2556325.2566237

Kim, J., Li, S., Cai, C. J., Gajos, K. Z., & Miller, R. C. (2014). *Leveraging video interaction data and content analysis to improve video learning*. Paper presented at the CHI 2014 Learning Innovation at Scale workshop, Toronto, Canada. Retrieved from https://dash.harvard.edu/handle/1/22719144

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

Kintsch, W., Kozminsky, E., Streby, W. J., McKoon, G., & Keenan, J. M. (1975). Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behaviour, 14*, 196-214.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). *Deconstructing disengagement: analyzing learner subpopulations in massive open online courses*. Paper presented at the Proceedings of the Third International Conference on Learning Analytics and Knowledge, Leuven, Belgium. http://dx.doi.org/10.1145/2460296.2460330

Klare, G. R. (1974–1975). Assessing readability. *Reading Research Quarterly, 10*, 62-102. http://dx.doi.org/10.2307/747086

Li, N., Kidzinski, L., Jermann, P., & Dillenbourg, P. (2015). *MOOC Video Interaction Patterns: What Do They Tell Us?* Paper presented at the Proceedings of the 10th European Conference on Technology Enhanced Learning Toledo,

Spain. https://dx.doi.org/10.1007/978-3-319-24258-3_15

Louwerse, M. M. (2001). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics, 12*, 291-315. http://dx.doi.org/10.1515/cogl.2002.005

Mayer, R. E. (2009). *Multimedia learning*. Cambridge, UK: Cambridge University Press.

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. http://dx.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, M.A: Cambridge University Press.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (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. https://dx.doi.org/10.1207/s1532690xci1401_1

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. http://dx.doi.org/10.1080/01638530902959943

McNamara, D. S., Raine, R., Roscoe, R., Crossley, S. A., Jackson, G. T., Dai, J., . . . Graesser, A. C. (2012). The Writing-Pal: Natural Language Algorithms to Support Intelligent Tutoring on Writing Strategies *Applied Natural Language Processing: Identification, Investigation and Resolution* (pp. 298-311). Hershey, PA, USA: IGI Global.

Miller, G. A. (1995). WordNet: a lexical database for English. *Commun. ACM, 38*(11), 39-41. http://dx.doi.org/10.1145/219717.219748

Millis, K. K., Golding, J. M., & Barker, G. (1995). Causal connectives increase inference generation. *Discourse Processes, 20*(1), 29-49. http://dx.doi.org/10.1080/01638539509544930

Mirriahi, N., & Vigentini, L. (2017). Analytics of Learner Video Use. In C. Lang, G. Siemens, A. Wise & D. Gašević (Eds.), *Handbook of Learning Analytics*. http://dx.doi.org/10.18608/hla17.022

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). LIWC2007: Linguistic inquiry and word count. [Computer software]

Pentimonti, J. M., Zucker, T. A., Justice, L. M., & Kaderavek, J. N. (2010). Informational text use in preschool classroom read-alouds. *The Reading Teacher, 63*, 656-665. https://dx.doi.org/10.1598/RT.63.8.4

Risko, E. F., Foulsham, T., Dawson, S., & Kingstone, A. (2013). The Collaborative Lecture Annotation System (CLAS): A New TOOL for Distributed Learning. *IEEE Transactions on Learning Technologies, 6*(1), 4-13. http://dx.doi.org/10.1109/TLT.2012.15

Rosé, C. P., & Ferschke, O. (2016). Technology Support for Discussion Based Learning: From Computer Supported Collaborative Learning to the Future of Massive Open Online Courses. *International Journal of Artificial Intelligence in Education, 26*(2), 660-678. http://dx.doi.org/10.1007/s40593-016-0107-y

Sinha, T., Jermann, P., Li, N., & Dillenbourg, P. (2014). *Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions*. Paper presented at the Empirical Methods in Natural Language Processing Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses, Doha, Qatar. https://arxiv.org/abs/1407.7131

Templin, M. (1957). *Certain language skills in children: Their development and interrelationships*. Minneapolis, MN: The University of Minnesota Press.

Turney, P. D., Neuman, Y., Assaf, D., & Cohen, Y. (2011). *Literal and metaphorical sense identification through concrete and abstract context*. Paper presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, United Kingdom.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic.

Varner, L. K., Jackson, G. T., Snow, E. L., & McNamara, D. S. (2013). Linguistic Content Analysis as a Tool for Improving Adaptive Instruction *Proceedings of the 16th Internation Conference on Artificial Intelligence in Education* (pp. 692-695). Berlin, Heidelberg: Springer Berlin Heidelberg.

Vega, B., Feng, S., Lehman, B., Graesser, A., & D'Mello, S. (2013). *Reading into the text: Investigating the influence of text complexity on cognitive engagement.* Paper presented at the Proceedings of the 6th International Conference on Educational Data Mining.

Wang, Y., & Baker, R. (2015). Content or platform: Why do students complete MOOCs? *MERLOT Journal of Online Learning and Teaching, 11*(1).

Zhang, D., Zhou, L., Briggs, R. O., & Nunamaker, J. F. (2006). Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. *Information and Management, 43*(1), 15-27. http://dx.doi.org/10.1016/j.im.2005.01.004

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