The Effectiveness of Enhanced Captions on L2 Collocation Learning: An Eye-tracking Study

WU YUMENG

Department of English Language and Literature

The Graduate School of
Sungshin Women's University

The Effectiveness of Enhanced Captions on L2 Collocation Learning: An Eye-tracking Study

A Dissertation
Submitted to the
Graduate School of Sungshin Women's University

in partial fulfillment of the requirements
for the degree of
Doctor of English Language and Literature

WU YUMENG

April, 2025

This is to certify that we have examined the
Doctoral Dissertation of
Wu Yumeng
Submitted to Department of English Language and Literature

Approved as to style and content:

Thesis Advisor ..... 고정민 .....

Committee Chairman ..... 윤태진 .....

Committee Member ..... 정소우 .....

Committee Member ..... 최현혜 .....

Committee Member ..... 이민진 .....

The Graduate School of
Sungshin Women's University

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Research Background

In the context of globalization, learning English has become an essential skill for many people around the world. As English continues to become a global lingua franca, the ability to communicate effectively in English is increasingly seen as critical to academic, professional and social success. However, learning English is not an easy task, especially for non-native English speakers who often face challenges in mastering the complexities of the language, such as vocabulary, grammar, pronunciation, and especially collocation (Laufer & Waldman, 2011).

In order to address these challenges and improve students' English proficiency, many schools and educational organizations are using multimedia resources, such as videos with English captions or subtitles as important teaching and learning tools. Videos provide an engaging and contextualized platform for learners to not only hear and see English used in real-life situations but also to improve their listening comprehension and vocabulary acquisition. Research has shown that the use of visual and auditory input in language learning can greatly support vocabulary retention and comprehension (Markham, Peter & McCarthy, 2001), especially when paired with effective instructional strategies such as captions.

In this context, it is important to distinguish between subtitles and captions as they serve different functions in language learning. Subtitles are usually text that translates or transcribes the spoken language in a video and are often used to help viewers who do not comprehend the video (Danan, 2004). Subtitles are usually limited to dialogue and may not include non-verbal sounds or environment noise. In contrast, captions are usually include not only spoken dialogue, but also descriptions of non-verbal audio elements such as sound

effects, background music and other relevant auditory cues (Huang & Eskey, 1999). The main purpose of captions is to help people with hearing impairments more fully understand audiovisual content. In the second language multimedia learning research, same-language captions (text in the target language accompanying audio in the same language) is commonly used to promote language comprehension, vocabulary acquisition, and text enhancement effects (Montero Perez, Van Den Noortgate, & Desmet, 2018; Rajendran & Mustafa, 2023).

Given this clear distinction, the on-screen text used in this study fits the definition of captions as it involves same-language transcription. Therefore, in line with well-established terminology and theoretical precision, this study uses the term “captions” throughout, which is consistent with previous multimedia L2 studies (e.g. Teng, 2021; Montero Perez, Peters, Clarebout, & Desmet, 2014).

Previous research has shown that captions can significantly facilitate second language acquisition (SLA). For example, a meta-analysis conducted by Montero Perez, Van Den Noortgate, and Desmet (2013) demonstrated that captioned videos can substantially enhance L2 learners’ listening comprehension and vocabulary acquisition. Similarly, Wang (2019) found that both L1 and L2 captioned television programmes contributed to students’ vocabulary learning and comprehension, with L2 captions being particularly beneficial for form recognition and vocabulary acquisition. Together, these studies highlight the efficacy of captions in facilitating target vocabulary acquisition without unnecessarily complicating the processing of non-verbal auditory cues.

In recent years, the concept of enhanced captions has gained attention as a more specialized tool for SLA. Enhanced captions are designed to go beyond the basic functions of traditional captions to include additional cues or modifications to aid comprehension and learning. Visual input enhancement may include colour-coded words, bold target expressions or pop-up style to draw attention to specific vocabulary or other linguistic features. These enhancements can help learners focus on key language items and promote deeper engagement

with the material. Research has shown that enhanced captions are more effective in promoting vocabulary acquisition and retention than non-enhanced captions by providing visual cues that link spoken and written language (Lee & Révész, 2020). In addition, enhanced captions can help learners understand more complex linguistic features such as collocations (eg., heavy rain), idiomatic expressions (eg., spill the beans) and other formulaic expressions that frequently occur in natural discourse, including discourse markers and fixed sentence stems (eg., you know what I mean) (Choi, 2023; Cintrón-Valentín & García-Amaya, 2021).

Perez, Van Den Noortgate, and Desmet (2013) conducted a meta-analysis of the effects of captioned and enhanced captioned videos in SLA and concluded that both traditional and enhanced captions improve listening comprehension, but enhanced captions tend to have a greater impact on vocabulary acquisition and retention. Research also found that enhanced captions increase learners' attention to collocations which in turn increases learner engagement and makes the learning process more focused and productive (Choi, 2017).

Despite the fact that studies have shown significant positive effects of enhanced captions on L2 learners' listening comprehension and vocabulary acquisition (Choi, 2017; Vu & Peter, 2023), there is still relatively little research on the effects of captions with enhanced target collocations, especially on the order of caption presentation. The literature focused on the general effects of captions, but there is a lack of research on the specific effects of the sequence of caption presentation, i.e., whether learners watch a captioned video with enhanced collocations before a non-captioned video, or a non-captioned video before a captioned video with enhanced collocations. For second language learners, it is often difficult to fully grasp the linguistic content of a video after watching it once, so videos are often shown multiple times to help students consolidate and deepen their memory.

1.2 Research Gap and Purpose

Previous studies do not provide sufficient answers to the question of whether the order of caption presentation affects learners' collocation learning and comprehension when they watch the same video multiple times. For example, watching a video without captions before watching a captioned video with enhanced target collocations may help learners to focus on listening comprehension and contextualization during the first viewing, while the presentation of captions with enhanced collocations during the second viewing may help learners to identify and reinforce the use of target expressions. Conversely, viewing a captioned video with textual enhancement first and then a non-captioned video may help learners to familiarize themselves with the collocations through the captions on the first exposure, and on the second viewing, it may help to deepen their understanding of these target collocations in the actual context.

Eye-tracking technology has become a valuable tool for exploring the cognitive processes involved in language learning. Eye-tracking technology enables researchers to track participants' visual attention in real-time, providing insights into how learners allocate their attention while watching videos. This technology has been used to study the impact of visual input on second language vocabulary learning, attention, and comprehension (Godfroid, Housen, & Boers, 2013). In the context of enhanced captions, eye-tracking studies are particularly valuable because they can provide detailed analysis of how learners interact with highlighted words, collocations, idiomatic expressions, and structural features such as syntactic patterns. However, the role of caption order in influencing attention and retention by using eye-tracking technology remains underexplored.

This gap in the literature is particularly important because the order of captions may affect learners' cognitive processing and attention, potentially impacting the effectiveness of video-based learning. By investigating this issue,

this study aims to fill a significant research gap. Using eye-tracking technology, this study will explore the how caption presentation order will influence second language (L2) collocation learning. Specifically, this study will figure out how the presentation order of the captions with enhanced target collocations—whether they are presented first or second during repeated viewings of the video—affect second language learners' collocation learning, comprehension, and attention. The study will examine two main conditions: one in which learners watch the captioned video with enhanced collocations first and then watch the same video without captions, and the other in which the order will be reversed—learners watch the video without captions first and then watch the captioned video with enhanced collocations.

1.3 Significance of the Study

The significance of this study is that it addresses an under-explored area in the existing literature regarding the order of captions in second language learning. While previous research has demonstrated the beneficial effects of captions in improving listening comprehension and collocation learning (Vanderplank, 2016), little attention has been paid to the specific order in which captions are presented during multiple viewings of a video. In a typical classroom setting, learners usually watch a video multiple times to enhance their comprehension and retention of language points. However, the question of whether the presence of captions first or last affects the learning process has still not been adequately addressed.

The results of this study will help us understand how captions order affects learners' cognitive processes such as attention allocation and cognitive load when learning second language content. By utilizing eye-tracking technology, this study aims to understand how learners allocate their visual attention when processing captions with enhanced target collocations and how this affects their attention to these collocations in real time. By doing so, a

more in-depth analysis of how learners interact with video-based linguistic input will be possible, thus providing more details on which the order of caption presentation maximizes learning outcomes.

Additionally, this study will contribute to the ongoing debate about the effectiveness of enhanced captions. While it has been shown that enhanced captions can greatly aid vocabulary retention (Perez et al., 2013), it remains unclear whether presenting enhanced captions at different stages of video viewing maximizes the learning experience. The present study aims to bridge this gap by investigating whether presenting captions with enhanced target collocations first or last affects collocation learning, comprehension, and retention. This will have practical implications for language teaching and learning, especially in multimedia-based learning environments where the use of captions can be easily adapted to optimize learning.

Besides, the results of the study could inform the design of more effective language learning materials. If research findings suggest that one sequence is more beneficial than another, then language educators can use these insights to adapt video content to English language instruction to improve students' learning. Understanding the optimal timing of caption presentation can help educators adapt the learning experience to better engage learners and improve their overall English proficiency.

In conclusion, the importance of this study is that it provides empirical evidence for a key aspect of video-based language learning—caption sequencing—that has practical applications in language teaching and multimedia content design. The findings of this study are expected to inform the development of more efficient and effective teaching strategies that combine visual and auditory input to enhance second language learning.

1.4 Thesis Structure

Chapter 1 introduces the background of this study, identifying the

significance of enhanced captions in second language (L2) learning, particularly in terms of linguistic learning and comprehension. The introduction part also describes the objectives and significance of the research, and discusses the gaps in the existing literature on the effects of caption presentation order on L2 learners. Chapter 2 is grounded in key theoretical perspectives, including incidental learning, attention, and awareness, which together inform the conceptual foundation of this study. The section also provides an in-depth review of the literature on second language acquisition (SLA), focusing on the use of captions, textual enhancement techniques, and eye-tracking technology in L2 learning. Chapter 3 is the methodology. This part explains the experimental setup, participants and target collocations selection, as well as the use of eye-tracking technology to analyze learners' eye-movements with eye-tracking measures. The chapter details the two experimental conditions—viewing the captions with enhanced collocations first, followed by no captions, and in reverse order—and the materials used (video clips, pretest, comprehension questions and immediate posttest). Chapter 3 also describes the data collection procedures and the statistical methods used to analyze the data. Chapter 4 presents the results of the study, focusing on the analysis of the eye-tracking data as well as the participants' comprehension and tests scores. The chapter reports the results of the study in terms of attention allocation in different caption conditions, highlighting how the order of caption presentation affects learners' visual attention, collocation learning and overall video comprehension. Statistical tests are used to assess the significance of the results and thereby provide a clear understanding of how the order of caption presentation affects learning outcomes. Chapter 5 explains the findings of Chapter 4 in the light of the research objectives. This section discusses the implications of the findings for theories of second language learning, particularly in relation to collocation learning and cognitive processing during video learning. It critically examines how the order of caption presentation affects learners' attention and collocation learning, and compares these results with those of previous studies. It also

explores the potential reasons behind the observed effects and considers the implications for language teaching, particularly in multimedia learning environments. The last chapter summarizes the main findings of the study and its implications for second language learning, particularly in terms of optimizing the use of captions in language learning materials. This part also reflects on the theoretical and practical application of the research, providing language educators and instructional designers with suggestions for effectively integrating captions into video learning. Finally, the chapter discusses the limitations of the study, including potential problems with sample size or experimental design, and makes recommendations for future research.

Chapter 2 Literature Review

2.1 Theoretical Framework: Incidental Learning, Attention, and Awareness

This study is grounded in several key theoretical constructs to explain how second language (L2) learners process and learn linguistic input, particularly in the context of learning collocations through captioned videos. Specifically, the framework of incidental learning, along with the cognitive constructs of attention and awareness, provides a foundational basis for explaining and understanding the internal mechanisms that govern L2 collocation learning in visually enhanced contexts such as captioned videos.

Incidental learning refers to acquisition of knowledge that occurs in situations that are not primarily intended for learning (Hulstijn, 2001). In second language acquisition (SLA) research, incidental learning of collocations typically occurs when learners are exposed to meaningful, context-rich input materials, even when they are not explicitly asked to focus on specific linguistic forms (Pellicer-Sánchez, 2016). Multimodal input environments, such as videos with captions, provide rich ground for incidental learning, as the target language form is naturally integrated into the authentic context. Textual enhancement further support the occurrence of incidental learning by visually highlighting the target items in the input while avoiding the direct intervention of explicit instruction (Lee & Révész, 2020). The enhanced captions used in this study aims to facilitate learners' incidental learning of the target collocations by increasing the perceived salience of the collocations.

However, incidental learning does not occur naturally; it is strongly modulated by the allocation of attention. Attention, often defined as the process of selective allocation of cognitive resources to specific stimuli (Gass, 1997; Robinson, 2003), plays a key role in the language learning process. According to

the Noticing Hypothesis (NH) proposed by Schmidt (1990), only input that is made aware can be transformed into possible resources for learning. Eye-tracking technology provides a powerful methodological tool for the objective measurement of learner attention by recording measures such as the number of visits, the total amount of fixation time during video viewing (Godfroid, 2020), revealing the allocation of learners' attention to different parts of the input, and there is an important link between these data and learning outcomes.

Closely related to attention is the concept of awareness, which refers to the conscious recognition of linguistic forms in the input by the learner (Schmidt, 1990; Tomlin & Villa, 1994). Awareness is considered a primary condition for transforming input into intake. The design goal of enhanced captions is precisely to stimulate learners' awareness of target collocations by enhancing visual salience. Although learners may not engage in explicit, and reflective learning processes, the higher level of awareness triggered by textual enhancement is expected to facilitate more effective collocation learning.

These theoretical foundations highlight the central role of attention and awareness in incidental language learning. Building on this, the present study further deepens related research by exploring how captions enhance visual input, guide learners' attention, and influence their collocation learning outcomes.

2.2 Captions and Textual Enhancement in L2 Collocation Learning

2.2.1 Collocation in L2 Learning

In modern linguistics, collocation refers to the fact that certain lexical items tend to co-occur more frequently in natural language use than would be predicted by grammar or meaning alone (Krishnamurthy, 2006). The concept of *collocation* was formally defined and popularized by British linguist John R.

Firth in the 1930s, although the term itself derives from the Latin word *collocare*, meaning "to place together". Firth(1957) distinguished *collocation* from the cognitive and semantic notions of word-meaning by treating it as a separate linguistic level within his framework for linguistic levels at which meaning emerges. According to Firth(1957), collocations represent deeper patterns of language use rather than being merely random or incidental word combinations. He argued that understanding these patterns is crucial to learning a language because it entails grasping combinations of words that native speakers frequently utter together, which often have meanings that cannot be inferred from the individual words. Firth's study established the foundation for later linguistic studies and approaches to teaching languages that emphasize the benefits of collocations for improving language proficiency and fluency. In Sinclair's work (1987), he introduced a statistical perspective to the concept of collocation in lexicography, defining collocates as "words which co-occur significantly with headwords." and regular or significant collocation as "lexical items occurring within five words... of the headword." This innovative approach marked the first incorporation of a quantitative measure to identify collocations in the field of dictionary compilation.

Richard Schmidt's Noticing Hypothesis (1990), highlighted the crucial role of awareness in language learning, positing that learners must consciously notice collocations in the input to effectively acquire them. This concept paved the way for the contributions of Michael Lewis (1993), which shifted the focus of language teaching towards "chunks" of language, including collocations, suggesting that fluency arises from the acquisition of these lexical combinations rather than from grammar alone. Benson, M Benson, E and Ilson (1997) classified collocations into two broad categories: grammatical collocations and lexical collocations. Grammatical collocations involve the combination of a subject word (usually a verb, adjective, or noun) with a grammatical structure (e.g., preposition or particle) (e.g., *interested in*, *dependent on*). On the other hand, lexical collocations involve the combination of content words, such as verb-noun

(*make a decision*), adjective-noun (*beautiful flower*), or noun-noun (*data analysis*) patterns. In 2005, Michael Hoey's Lexical Priming theory offered a significant advancement in understanding how collocations are acquired in language learning. Hoey suggested that exposure to language shapes subconscious "primings" which make certain word combinations feel more natural than others. According to this theory, repeated exposure to specific collocations in various contexts primes learners to use these combinations more frequently and fluently in their own speech and writing. This priming effect underscored the importance of extensive reading and listening in language education, as it enhances the likelihood that learners will acquire collocations not through deliberate study but through repeated, meaningful exposure. This theory has been instrumental in explaining why some collocations become solidified in learners' minds and has influenced teaching practices to incorporate more authentic language use in educational settings.

Over the past decade, empirical research on second language (L2) collocation learning has provided valuable insights into the effectiveness of explicit teaching and implicit learning as instructional methods. Explicit teaching has emerged as the most effective approach, particularly in the learning and retention of complex and low-frequency collocations. For instance, Askari (2024) emphasized the necessity of direct instruction in EFL contexts, arguing that explicit teaching provides a structured framework essential for mastering collocations beyond basic lexical combinations, as well as highlighting the linguistic, theoretical, and pedagogical importance of collocation instruction, particularly in relation to the four language skills of listening, speaking, reading, and writing. Similarly, Huang, Abdul Samat, and Haladin (2024) further reinforced this position. They demonstrated that explicit teaching outperformed implicit strategies when learners were exposed to diverse collocational patterns. These studies highlight the advantages of structured guidance in collocation learning, particularly for L2 learners at lower proficiency levels or when dealing with less frequent collocations.

By contrast, research on implicit learning has shown mixed results. Li's (2023) study investigated the effectiveness of data-driven learning (DDL) as a form of implicit instruction in enhancing the accuracy of collocation use among L2 learners in writing tasks and supported the effectiveness of implicit instruction through DDL learning, particularly when paired with indirect corrective feedback. Implicit instruction was operationalized through exposure to authentic language use and guided corpus consultation, allowing learners to inductively identify and correct collocational inaccuracies. Building on the earlier work of Sonbul and Schmitt (2013), Toomer and Elgort (2019) aimed to extend the understanding of how implicit instruction influences collocation acquisition and retention over time. Implicit instruction was achieved through extensive exposure to collocations embedded in meaningful language tasks rather than direct rule-based teaching. The findings indicated that while implicit learning led to some improvement in the recognition and comprehension of frequently encountered collocations, it was less effective in promoting productive collocation use compared to explicit instruction.

The findings align with the growing consensus in second language (L2) learning research advocating for a balanced pedagogical approach that integrates both implicit exposure and explicit metalinguistic feedback for optimal collocation learning.

2.2.2 The Role of Captions in L2 Learning

Captions have been widely investigated in second language (L2) learning as a tool to facilitate vocabulary learning. Paivio's Dual Coding Theory (1986), proposed that human cognition involves two cognitive channels for processing and representing information: a verbal channel for linguistic input and a non-verbal (visual) channel for imagery-based input. When learners are exposed to captioned audiovisual material, they receive both auditory input from spoken language and visual input from the captions. This dual representation

strengthens memory traces and supports deeper processing, particularly for complex language units such as collocations. Paivio (1986) suggested that presenting information in both formats concurrently can facilitate learning because it provides multiple pathways for retrieving information.

Recent empirical studies have extensively explored the role of captioned videos as a pedagogical tool in facilitating L2 vocabulary and collocation learning, with findings consistently supporting Paivio's Dual Coding Theory (DCT) (1986). Consistent with this, Teng (2021) found that learners who watched captioned videos made much greater gains in incidental vocabulary learning compared to learners who watched non-captioned content. The visual reinforcement of spoken language by captions enabled learners to make stronger form-meaning connections, which in turn facilitated vocabulary retention. Similarly, Teng and Cui (2025) in a study comparing the effects of different caption types on vocabulary and collocation learning showed that full captions were the most effective in facilitating vocabulary and collocation learning, and that learners' knowledge of vocabulary and working memory capacity had a significant effect on learning outcomes. Rajendran and Mustafa (2023), on the other hand, emphasized that graded captioned videos are good facilitators of vocabulary and collocation learning for low level learners, and they pointed out that the multimodal input helps learners to process the target expression more deeply.

These findings further solidify the role of captions as a valuable instructional tool in L2 learning, as they provide a rich multimodal learning environment that aligns with cognitive theories of language processing and dual-channel encoding.

2.2.3 Textual Enhancement: Definitions and Benefits

Textual Enhancement (TE) refers to the deliberate manipulation of written input in ways that make specific linguistic features more salient and noticeable

to second language (L2) learners. This is achieved by modifying the visual presentation of the target linguistic elements through various means such as bolding (e.g., Peters 2009, 2012; Sonbul and Schmitt 2013; Toomer and Elgort 2019), underlining (e.g., Boers et al. 2014; Majuddin et al. 2021; Puimège et al. 2023; Szudarski and Carter 2016), or coloring (Jung et al. 2022, 2024). The goal of TE is to increase learners' awareness of certain features of the language (e.g., words, grammar, and collocations) within the input, thereby facilitating their learning and comprehension.

Textual Enhancement is based on the theoretical foundation of Schmidt's Noticing Hypothesis (1990), which argues that for language features to be successfully learned, they must first be consciously noticed by the learner. By drawing learners' attention to these forms within context-rich input, TE helps foster the connection between form and meaning, which is essential for language learning. When learners notice an enhanced form (e.g., a collocation, a grammatical structure) in a meaningful context, they are more likely to understand "how it is used"—that is, the actual linguistic function to which the form corresponds and thus establish a "form-meaning" connection in the brain. Additionally, TE aligns with Paivio's Dual Coding Theory (1986), which posits that language learners process information more effectively when it is presented through both verbal and visual channels simultaneously to enhance memory and recall.

Studies have shown that this heightened attention helps learners better acquire both vocabulary and collocation (Goudarzi Z., & Moini, M. R. 2012; Fazlali, B., & Shahini, A. 2019). Furthermore, textual enhancement has been found to support the learning of grammatical structures by making them more salient in the input, which contributes to greater grammatical awareness and improved language production. For instance, Abbasian and Yekani's study (2014) contributed to the growing body of literature on the effectiveness of textual enhancement as a tool in L2 grammar learning. It highlighted that visual enhancement of target forms in texts can help learners notice and internalize

grammar structures without the need for explicit instruction.

In addition to vocabulary and grammar, textual enhancement facilitates comprehension and retention by drawing learners' focus to key linguistic elements, thus enhancing their ability to retain and recall information in the long term. Peters (2012) provided compelling evidence that typographic salience (e.g., bolding and underlining) significantly improves the recall of formulaic sequences in L2 learning contexts. However, explicit instructional focus without visual modifications may not be sufficient for maximizing language retention. These findings reinforced the role of input enhancement techniques in supporting L2 learning, particularly for complex linguistic forms like formulaic sequences (Han, Park, & Combs, 2008; Lee & Huang, 2008; Simard, 2009).

Another key benefit is its ability to foster incidental learning by highlighting target features within natural, context-rich input, which allows learners to acquire language forms without explicit instruction. Jung and Lee's (2024) study investigated the role of incidental learning in second language (L2) collocation learning by examining the effects of reading-while-listening (RWL) activities enhanced with synchronized textual input modifications. The results showed that learners in the textually enhanced RWL group demonstrated significantly higher retention of collocations compared to the non-enhanced group. This finding suggested that collocations were acquired incidentally, without the need for explicit metalinguistic instruction. The increased retention of collocations in the enhanced group supports the theory that incidental learning can occur during meaningful exposure to language in context.

Based on the previous studies, textual enhancement provides significant support for incidental language learning, making it a versatile and effective strategy for enhancing L2 learning across a range of linguistic domains, including vocabulary, collocation, and grammar.

2.2.4 Different Views on Textual Enhancement in L2 Learning

Vu and Peters (2022, 2023) explored whether textual enhancement could promote incidental learning of collocations in various input modes, such as reading-only, reading-while-listening, and reading-while-listening combined with textual enhancement (i.e., underlining). In these studies, L1 Vietnamese speakers participated in long-term programs where they read graded readers containing 32 target collocations under a counterbalanced design. The findings indicated that textual enhancement significantly facilitated the learning of target collocations, regardless of the input mode, and reinforced the beneficial impact of textual enhancement on incidental collocation learning. Similarly, Jung and Lee (2024) investigated the potential of textual enhancement in promoting incidental collocation learning, focusing on whether the timing of textual enhancement—either static (consistently visible throughout the video) or synchronized with audio (appearing at the moment of the target word)—affected its efficacy. The results demonstrated that both static and synchronized textual enhancement (coloring) promoted receptive knowledge of target collocations, with synchronized coloring further boosting semantic processing of the collocations.

However, Montero et al. (2014) presented a contrasting viewpoint in their research, where they examined the effectiveness of three types of captioned videos, varying in the amount of text and the salience of lexical items, for content comprehension and incidental vocabulary learning in L2 French. The study found that the type of captioning (full captioning, keyword captioning, or full captioning with highlighted keywords) did not significantly impact comprehension scores, indicating that the salience of lexical items in captions did not enhance content understanding. In terms of vocabulary learning, the study showed that captioned groups outperformed the control group on form recognition and clip association tests. However, no significant differences were observed between the different types of captioning, which suggests that the addition of lexical enhancement through keyword highlighting did not lead to

greater vocabulary gains compared to full captioning without enhanced lexical items. This finding aligns with the results of Majuddin et al. (2021), who investigated the effects of textual enhancement (via bolding and underlining) and repeated viewing of captioned videos on the learning of target multiword units. While repeated viewing was found to facilitate learning of the target units, textual enhancement had a negligible impact on learning outcomes, and in some cases, it even diminished the positive effects of captions on content comprehension.

These contrasting findings (Vu and Peters, 2022; Jung and Lee, 2024; Montero et al., 2014) underscore the complexity of the effects of textual enhancement and suggest that its efficacy may be influenced by several factors. These include the type of target language unit (e.g., collocations vs. individual words), the type of captions used in the videos (e.g., full, keyword, or full captions with enhanced keywords) and the timing of the presentation of the enhancement (e.g., whether it is static or synchronized with audio).

2.3 Eye-Tracking in Language Learning Research

2.3.1 Eye-Tracking Methodology and Language Processing

Eye-tracking involves using specialized equipment to record the movements and fixation duration of the eyes as they engage with various stimuli. Using near-infrared light and high-resolution cameras, an eye tracker records moment-to-moment eye movements and measures what a learner is looking at and how long that learner gazes at a region (Holmqvist, Nyström, Andersson, Dewhurst, Jarodzka, & van de Weijer, 2011). Thus, the eye tracker has been considered an objective, direct, and real-time measure of visual attention. This technology has proven to be particularly valuable for studying how individuals process and interpret written and spoken language, as well as offering a non-invasive window into cognitive processes such as attention, memory, and language comprehension.

Eye-tracking technology has made valuable contributions to language processing research due to its ability to measure real-time cognitive processing during tasks such as listening, reading and viewing multimedia.

In the context of second language (L2) learning, auditory processing plays a key role in developing listening comprehension skills. Eye-tracking studies have demonstrated that L2 learners are more likely to rely on visual input (captions, pictures, or written cues) when they struggle to understand spoken language. For example, Kho, Aryadoust, and Foo (2023) explored how listeners process auditory input during listening assessments, specifically examining the role of the keyword-matching strategy, an approach in which listeners try to match words they hear to keywords in the written modality (test items), using eye-tracking technology. The primary goal was to understand how the participants utilized visual information (e.g., written keywords) to aid their auditory processing of spoken content. The eye-tracking results demonstrated a direct relationship between gaze patterns and auditory processing. Specifically, listeners' eye movements were synchronized with the unfolding speech. Longer fixations on the spoken content coincided with critical moments in the auditory stimuli. It means that participants were processing and verifying the auditory information as it was being presented. Eye movements showed that learners were able to integrate auditory and visual input effectively, which can enhance the efficiency of auditory processing during language assessments.

As for the reading of visual world processing, eye-tracking allows researchers to monitor where, when, and for how long a reader fixates on particular words or regions of text, then offer valuable data on reading comprehension, lexical processing, and sentence parsing. Research has shown that longer fixations tend to occur on unfamiliar or complex words, which signals that readers are investing more cognitive resources into understanding those terms (Rayner, 2009). Another important aspect is the time-course of reading, which means how reading develops over time, including the timing and length of fixations on specific words. The studies (Zagar, D. et al 1997; Zhang

et al 2021) often investigated first-pass reading (when a word is encountered for the first time) or second-pass reading (when a word or phrase is revisited for clarification or deeper processing). In addition to word-level processing, sentence parsing and syntactic processing are also illuminated by eye-tracking. For instance, when readers encounter syntactically complex sentences, eye-tracking can show patterns of increased gaze duration, which indicates the difficulty in processing sentence structure or ambiguity (Demberg & Keller 2008; Mertzen, D et al 2023). By observing how readers navigate linguistic and visual input, adjust their focus, and process meaning across time, eye-tracking offers a precise window into cognitive processing during reading.

Other research has examined how visual and auditory linguistic stimuli are processed in integrated activities. Xie's (2019) study examined the potential benefits of coordinating visual and auditory cueing in multimedia learning. The researchers conducted three eye-tracking experiments to investigate whether providing coordinated visual and auditory cues (where key elements were highlighted in the graphic and emphasized in the narration) could improve learners' attention to relevant information and enhance their learning outcomes, compared to providing no cues, only visual cues, or only auditory cues. The findings showed that coordinated dual cues led to better performance on posttests and greater attention to the relevant parts of the graphics, compared to the other cueing conditions. These results extend the signaling principle in multimedia learning, and suggested that coordinating visual and auditory cues can be an effective way to guide learners' cognitive processing and promote meaningful learning. Studies have looked at the processing of visual and auditory linguistic stimuli in dual code activities.

2.3.2 Eye-Tracking and Acquisition in L2 Learning

Many studies use eye-tracking to observe how learners interact with new vocabulary during reading tasks. For instance, Godfroid, Ahn, Choi, Ballard, Cui,

Johnston, Lee, and Sarkar (2018) demonstrated that longer fixations on novel words correlated with better vocabulary retention, and reveled that increased attention enhances learning. This method has shown that learners tend to spend more time on unfamiliar words, and those linguistic forms that receive increased attention are more likely to be remembered. In listening tasks, eye-tracking has been used to examine how learners allocate their visual attention when processing spoken vocabulary. For example, Conklin and Pellicer-Sánchez (2016) used a visual world paradigm to track eye movements as learners listened to sentences containing target vocabulary. Their study found that learners' gaze was drawn to images that matched the spoken vocabulary, indicating that visual context and auditory input together enhance the learning of new words through multimodal integration.

Eye-tracking research also extends to the study of collocations in L2 learning. By examining eye movements, researchers can assess how learners process and learn fixed expressions or common word pairs. For example, Sánchez's (2022) investigated the processing and acquisition of novel words and their collocates (i.e., words that frequently co-occur with other words) through reading, as well as the impact of exposure frequency on this process. The findings demonstrated that participants gained knowledge not only of the form and meaning of pseudowords but also of their collocates. Eye movement analysis revealed a significant effect of exposure frequency on the processing of novel collocations for both first and second language readers, with reading times decreasing as exposure increased. While another study by Vilkaitė and Schmitt (2019) investigated the processing of collocations in a second language (L2), with a particular focus on whether processing benefits extend beyond adjacent collocations to non-adjacent ones. The study involved L2 English learners reading sentences that contained both adjacent and non-adjacent collocations. The results revealed that L2 readers processed adjacent collocations more quickly and accurately than non-adjacent collocations, which shows that the proximity of collocating words plays a crucial role in the ease of processing.

This study also highlighted the potential for processing benefits to transfer across different types of collocational structures, but the benefits were more pronounced for adjacent collocations, likely due to their stronger associative ties. The authors concluded that while collocational processing can extend to non-adjacent forms, the processing benefits are stronger when collocations are adjacent. The study by Li (2022) also aligns with the findings of Vilkaitė and Schmitt (2019) in terms of investigating exposure frequency and its effects on processing.

2.3.3 Eye-Tracking and Visual Input Enhancement

In recent years, eye-tracking technology has been increasingly used to examine the effectiveness of captions—both standard and enhanced—in attracting learners' attention to target expressions and its subsequent impact on learning outcomes. Eye-tracking provides valuable insights into how learners process written language while interacting with audiovisual content and offering a more detailed understanding of the cognitive mechanisms involved in second language (L2) learning.

Park et al. (2012) conducted a study on visual input enhancement, attention, and grammar learning in reading comprehension, using eye movement data to show how visual salience impacts the processing of grammatical structures (gerund and to-infinitives). Similarly, Kim et al. (2023) explored the effects of announcing a vocabulary test before reading a glossed text, by using eye-tracking to analyze reading behaviors and vocabulary learning, emphasizing the relationship between learner attention and retention.

In the context of multimedia and digital input, Choi (2023) specifically examined the effect of visual saliency in captioned digital videos on the learning of English collocations. The study highlighted the influence of visual cues in increasing learners' attention to target vocabulary, and found that enhanced captions can facilitate the processing and retention of collocations. In line with

this, Lee and Révész (2020) focused on the role of captions and textual enhancement in multimodal input tasks, and explored how these factors contribute to the promotion of grammatical development—specifically, the learning of the present perfect versus past simple tense distinctions—in L2 learners. Their study emphasized how multimodal learning environments, with a combination of visual and textual input, can support linguistic development such as tense-aspect distinction by guiding attention and processing.

Further exploring the influence of textual enhancement, Lee and Jung (2024) examined the effects of textual enhancement and task manipulation on L2 learners' attentional processes and grammatical knowledge development, which provides insights into how tasks designed to highlight textual features can influence learner focus and cognitive processing. Likewise, Puimège et al. (2023) focused on the learning of multiword units through textually enhanced audiovisual input, using eye-tracking to show that enhanced input significantly influences learners' attention and learning outcomes, particularly in the context of fixed collocations. Then, Puimège et al. (2024) continued this line of inquiry by investigating the effects of typographic enhancement on L2 collocation processing and learning from reading. Their eye-tracking study emphasized how typographic features such as bolding and underlining affect learners' attention to collocations, and showed that these visual cues help guide learners' focus to important lexical items, enhancing both processing and retention.

Overall, while the methods and areas of focus vary, the common thread remains the importance of using traditional visual cues, that is, enhancement (e.g., bolding or color coding), to enhance language learners' attention and subsequent language learning.

2.4 Summary

The literature review reveals several key findings regarding the effectiveness of captions, and textual enhancement through eye-tracking

technology in second language (L2) learning, particularly in the context of collocations. Captions have been widely recognized as an effective tool in L2 learning that enhances video comprehension and facilitates linguistic forms learning by reinforcing auditory input in the form of synchronized text. Textual enhancement, which highlight key or collocations, have shown promise in further improving collocation learning by making target expressions more salient and facilitating their contextual understanding. Eye-tracking studies, which track learner's attention, indicate that learners tend to allocate more attention to captions than to the spoken input, which can enhance collocation learning (Park et al., 2012; Lee and Jung, 2024).

Despite these advancements, notable gaps remain in the literature, particularly regarding the order of caption presentation (captions with enhanced collocations first vs. last) and its potential effects on video comprehension and collocation learning. Few studies have explored how the sequencing of caption presentation impacts learner engagement and learning outcomes, especially with respect to collocations. Although some research has examined the general benefits of enhanced captions in collocation learning, very limited attention has been paid to how the order of caption exposure (e.g., captions with enhanced collocations first vs. non-captions first) may affect the depth of processing or retention of collocations. Given that collocations often requires more complex contextual support and repeated exposure to be effectively acquired, so whether the sequence of caption presentation affects learners' attention and processing of collocations has not been systematically explored. Previous studies have generally not controlled for the variable of caption order, and have less often incorporated process data (e.g., eye-tracking) to reveal cognitive processes. These limitations highlight the need for further empirical research to clarify the role of the order of captions with enhanced collocations in L2 learning.

2.5 Research Questions

Based on the above findings, this study, guided by the theoretical frameworks of incidental learning, attention and awareness, aims to investigate how the order of captions with enhanced target collocations affects visual attention allocation and collocation learning of second language learners. Specifically, by combining eye-tracking technology with captioned video containing enhanced collocations input, this study systematically examines the cognitive processing and learning effectiveness of learners under different caption presentation sequences, in order to fill the gaps in existing research. The findings will contribute to the understanding of how caption presentation and attention influence L2 learning, particularly in the collocation learning, and will offer practical implications for designing more effective L2 instructional materials. To achieve this, the study formulates the following research questions:

1. To what extent does the sequence in which EFL learners watch English-captioned and non-captioned videos (watch English-captioned videos with enhanced collocations first or later) influence their immediate posttest performance on enhanced target collocations and their video comprehension scores?

2. To what extent does the order of captioned and non-captioned video viewing result in distinct patterns in learners' eye-tracking measures (fixation counts, number of visits, and total fixation duration) to enhanced target collocation areas and to what extent are these three three eye-tracking measures interrelated?

3. To what extent are eye-tracking measures (fixation counts, number of visits, and total fixation duration) correlated with immediate posttest scores, and to what extent do they collectively predict EFL learners' final performance on enhanced target collocations?

Chapter 3 Research Methodology

3.1 Overall Design

This study employed a between-subjects experimental design to investigate the effect of enhanced English captions on second language (L2) learners' comprehension and learning outcomes, with a specific focus on target collocations. Sixty participants were randomly assigned to one of two experimental groups, which differed in the order of video viewing and the presence of captions. Group one (n=30) viewed a video with enhanced captions (containing target collocations) first, followed by a second video without captions (EC+NC Group), while another group (n=30) viewed the same videos in reverse order, starting with the video without captions and followed by the version of captions with enhanced target collocations (NC+EC Group).

The sequence of video viewing for each group was critical to understanding how the order of caption exposure (captions with enhanced collocations first vs. last) might influence both attention and learning outcomes. By systematically varying the video conditions between groups, the study aims to determine whether the presentation of captions with enhanced collocations early or late in the video have an effect on participants' fixation counts, number of visits, and total fixation duration, as well as their ability to understand the target collocations.

3.2 Participants

A total of sixty university students (57 females, 3 males) in Sungshin Women's University who used English as their foreign language (L2) were recruited as participants. Their ages ranged from 18 to 42 (Mean=23, *SD*=4.783). They were from various majors, including English, Korean, German, Japanese,

Arts, Accounting, Business Administration, Drama and Film, as well as Media Communication. While most of the participants had no experience living in an English-speaking country, 14 students reported that they had lived in the United States, Canada and Malaysia from two months to eight years. There are 39 Koreans, 19 Chinese and 2 Mongolians agreed to participate in the study. Among them, all Korean students have been studying English as a foreign language for approximately 10 years, and 26 of them haven taken the TOEIC, IELTS and TOEFL tests. The Chinese participants had completed the College English program in China. Among them, four non-English majors had passed the College English Test Band 4 (CET-4), which is a graduation requirement for non-English majors. Three participants had also passed the College English Test Band 6 (CET-6), indicating a higher level of English proficiency. In addition, three other participants, who majored in English, had passed the Test for English Majors Band 8 (TEM-8), a professional exam designed for English majors in Chinese universities. The other two Mongolian students began to learn English in the junior high school, later than Korean and Chinese students, and one of them scored 6.5 in the IELTS test.

In order to measure participants' differential English proficiency, the EF Standard English Test (EF SET) was used (*https://www.efset.org/*). This test is a free, online English proficiency assessment developed by EF Education First. It offers there major test formats: a 90-minute full skills test, a 50-minute test and a 15-minute quick test. Due to time constraints and to minimize participants' boredom, the 15-minute test was selected for this study. Although shorter, this version still provides a reliable estimate of participants' general English proficiency. The 15-minute test included reading (grammar and vocabulary) and listening to assess their English ability before the experiment begins, and the two parts contained about 10 comprehension multiple-choice questions with four options respectively. The score was converted into the level from Common European Framework of Reference for Languages (CEFR). Although it was created exclusively for any European language, the CEFR is

now a global standard framework for language proficiency, particularly in European languages like English. The EF SET is in line with the six levels of foreign language proficiency established by the CEFR ranging from A1 (Beginner), A2 (Elementary), B1 (Intermediate), B2 (Upper Intermediate), C1 (Advanced), to C2 (Proficient). This 15-minute online test provides an overall score (ranging from 0 to 100) and reports CEFR proficiency levels in the form of broad bands (e.g., A1-A2, B1-B2), rather than specific levels. From Table 1, it can be inferred that all participants are English learners whose English proficiency is from beginner to upper intermediate (from level A1-A2 to B1-B2).

Table 1 EF SET scores and corresponding CEFR proficiency bands of participants

Participant	EF EST Score	CEFR Band Level
1	75	B1-B2
2	60	A1-A2
3	55	A1-A2
4	80	B1-B2
5	80	B1-B2
6	50	A1-A2
7	50	A1-A2
8	75	B1-B2
9	55	A1-A2
10	80	B1-B2
11	50	A1-A2
12	50	A1-A2
13	65	B1-B2
14	55	A1-A2
15	55	A1-A2
16	70	B1-B2
17	50	A1-A2
18	55	A1-A2
19	50	A1-A2
20	70	B1-B2
21	50	A1-A2
22	55	A1-A2
23	55	A1-A2
24	50	A1-A2
25	55	A1-A2
26	60	A1-A2
27	50	A1-A2
28	50	A1-A2
29	60	A1-A2
30	75	B1-B2
31	85	B1-B2
32	55	A1-A2
33	65	B1-B2
34	55	A1-A2

35	55	A1-A2
36	55	A1-A2
37	55	A1-A2
38	50	A1-A2
39	60	A1-A2
40	50	A1-A2
41	80	B1-B2
42	70	B1-B2
43	50	A1-A2
44	65	B1-B2
45	50	A1-A2
46	55	A1-A2
47	55	A1-A2
48	70	B1-B2
49	65	B1-B2
50	55	A1-A2
51	65	B1-B2
52	50	A1-A2
53	55	A1-A2
54	50	A1-A2
55	65	B1-B2
56	50	A1-A2
57	50	A1-A2
58	70	B1-B2
59	50	A1-A2
60	60	A1-A2

Table 2 shows that the average English proficiency score was 59.00 (SD=11.477) for the EC+NC group and 57.83 (SD=10.803) for the NC+EC group.

Table 2 Descriptive statistics for EF SET scores between the two groups

	Group	N	Mean	Std. Deviation	Std. Error Mean
EF SET scores	EC+NC	30	59.00	11.477	2.095
	NC+EC	30	57.83	10.803	1.972

Then, an independent samples *t*-test (Table 3) was run to examine group differences in English proficiency.

Table 3 *T*-test results for EF SET scores between the two groups

		Levene's Test for Equality of Variances		*t*-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
EF SET scores	Equal variances assumed	.413	.523	.405	58	.687	1.167	2.878

The results revealed no significant difference between the two groups, $t(58)=0.405$, $p=.687$, indicating comparable English proficiency levels across the two groups.

3.3 Materials

3.3.1 Selected Videos

Two video episodes from an American documentary television series *Explained* that can be watched on the streaming service *Netflix* were used as stimuli in the eye-tracking experiment. One video is entitled with *Why Women Are Paid Less* which discusses the gender pay gap and the rights of women in the workplace, while the other *The World's Water crisis* talks about an examination of the water crisis around the world, including ideas on how to protect clean water and how to ensure that developing countries have access to safe drinking water. Each episode is voiced by a different guest narrator and both last 18 minutes. In order to increase the efficiency of the experiment and maintain the enthusiasm of the participants for the study, 4 minute 57 seconds clip from *Why Women Are Paid Less* and 4 minute 7 seconds clip from *The World's Water crisis* were selected. In addition, the two selected clips level were comparable in terms of the density and complexity of pictorial content, the level of correlation between visual information (image) and verbal information (narration), as well as the speech rate, which were 162.22 words per minute (803 words and 66 sentences), and 161.05 words per minute (663 words and 59

sentences) respectively. The Flesh - Kincaid Grade Level was 6.98 of *Why Women Are Paid Less* and 6.48 of *The World's Water crisis*, meaning that the comprehensibility level of the English captions from the clips approximated that of native English speakers who have received eight to nine years of schooling. These indices reveal that the linguistic difficulty level of the two videos were appropriate for the participants.

In order to answer the research questions, the researcher created two conditions for each video clip. One is the English captions, the other is the English captions with the enhanced target collocations. As shown in Figure 1, the target collocations were made visually salient only in the enhanced version. Unlike Lee and Révész (2018) who used boldface font to enhance target grammatical units, color font was utilized in this study because using boldface font may slightly alter the size of the regions of interest.

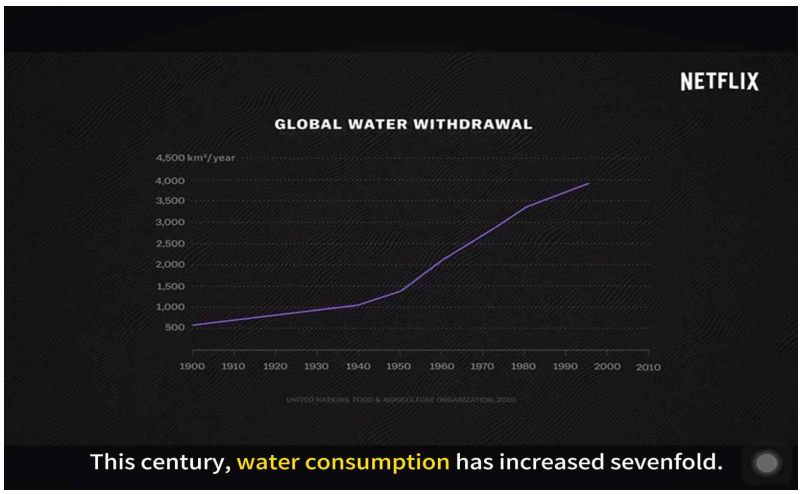

Figure 1 Enhanced target collocation

For English captions with the enhanced collocation version, full English captions were added by using *Jianying*, which is a professional desktop editing software in China.

3.3.2 Target Collocation Selection

Among the English captions, eight collocations in total were chosen as the target expressions. Four collocations from the video *Why Women Are Paid Less* (*hover around, a slew of, primary caregiver,* and *the trajectories of*) and another four items from *The World's Water crisis* (*water consumption, embedded in, common ingredient,* and *abide by*).

According to the classification proposed by Benson et al. (1997), the eight target collocations selected for this study include both grammatical and lexical collocations. Grammatical collocations usually consist of combinations of content words with prepositions, particles, or other grammatical structures, whereas lexical collocations consist mainly of collocations between content words, such as adjective-noun or noun-noun combinations. In the collocations used in this study, five of target collocations (*hover around, a slew of, the trajectories of, embedded in,* and *abide by*) were grammatical, while the rest of three (*primary caregiver, water consumption, common ingredient*) were lexcial. This mixture reflects the true distribution of collocation patterns in natural language use, and helps to examine more comprehensively the processing characteristics of different structural types of collocations under enhanced captioning conditions.

The target collocations used in this study were selected from the Corpus of Contemporary American English (COCA) to ensure that they represent authentic, and contextually appropriate language use. In order to show the distribution of each collocation in real language use in more detail, Table 4 shows their frequency distribution in nine COCA corpora (such as spoken, fiction, academic, etc.). According to the table, the frequencies of the target collocations selected for this study in the COCA corpus ranges from 61 to 5,068, covering different frequency bands with high-, and low-frequency categorizes. Based on the research of Nation (2001) and Schmitt and Schmitt (2014), five of these collocations (*hover around, primary caregiver, the trajectories of, water consumption,* and *common ingredient*) are low-frequency items (frequency<500),

and three collocations (*a slew of, embedded in, abide by*) belong to high-frequency items (frequency>2,000). High-frequency words usually refer to the most common 2,000-3,000 word families, which are more familiar to learners and less difficult to acquire, whereas low-frequency words, which have less exposure to real-world input, are more challenging for learners and more difficult to acquire and retain (Nation, 2006; Schmitt & Schmitt, 2014).

Table 4 Frequency of target collocations across different registers in the COCA Corpus

Target Collocation	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD
hover around	215	28	26	9	11	30	55	41	15
a slew of	2099	361	311	42	126	109	595	501	54
primary caregiver	216	19	22	14	19	6	40	27	69
the trajectories of	94	3	14	0	1	6	15	4	51
water consumption	260	38	37	4	5	1	54	17	104
embedded in	5068	502	623	186	201	478	821	325	1932
common ingredient	61	6	5	3	5	2	24	5	11
abide by	2631	453	437	145	509	99	221	411	356

Note. Frequencies indicate raw occurrence counts of each target collocation across different registers in the Corpus of Contemporary American English (COCA). Registers include: BLOG (blogs), WEB (webpages), TV/M (television and movies), SPOK (spoken), FIC (fiction), MAG (magazines), NEWS (news reports), and ACAD (academic texts).

It should be noted that Nation (2001) argues that very low-frequency words usually refer to vocabulary with fewer than 10 occurrences in a large corpus, and should be acquired mainly through extensive reading and contextual inferences rather than direct instruction. Although several low-frequency collocations in this study have relatively low frequencies, none fall into the "very low-frequency" category; instead, they reflect the types of collocation that L2 learners may realistically encounter but are likely to find unfamiliar in authentic contexts. By consciously covering low-frequency and high-frequency collocations, the present study is able to systematically examine the effects of

input enhancement and presentation order of captions on collocation learning at different word frequency bands, which more realistically reflects the collocation diversity that L2 learners are exposed to in actual language environments, and also increases the pedagogical relevance of the findings (Nation, 2001).

Table 4 also shows that high-frequency collocations (such as *embedded in* and *abide by*) are widely distributed across multiple registers, while low-frequency collocations (such as *common ingredient*) appear infrequently, especially in written registers.

Besides, five of them were presented in video once, while other three words—presented as single words without accompanying collocations—appeared more than once: the words *trajectory* appeared three times, while *embedded* and *ingredient* were displayed twice. Each video scene which featured a single-line caption presented at the bottom of the screen, using a proportional 30-point Arial font. The placement of the target collocations within the caption was not fixed; rather, they appeared in accordance with the flow of the narration.

In selecting the target collocations, this study used corpus frequency as the main criterion to ensure the authenticity and contextual relevance of the selected collocations. However, it is recognized that there are differences in the perceptual salience of different collocations, mainly in terms of phonological features such as the number of phonemes and syllable structure (Goldschneider & DeKeyser, 2001). For example, the collocation *a slew of* has a shorter and simpler phonological structure, whereas another collocation *the trajectories of* is longer and more complex, which may affect learners' attention allocation. Although phonetic complexity is not a controllable variable in the selection of collocations, efforts were made to reduce potential perceptual differences by presenting all collocations in a natural video context and applying consistent enhancement techniques to each collocation.

3.3.3 Collocation Pretest

The participants were required to take a paper pretest before the experiment. The pretest is consisted of three sections. The first section contained 20 collocations in total, including two target expressions and 18 distractors. For each item, subjects were asked to choose the most accurate meaning of the collocation from three multiple-choice options. The second part was a fill-in-the-blank task consisting of 15 sentences. Participants were required to complete each blank with the most appropriate collocation selected from the word bank containing three target expression and 12 distractors. The last step involved choosing the correct prepositional collocation from three multiple-choice options, which included three target expressions and 12 distractors. To sum up, there are eight target collocations which are enhanced English captions in the video and 42 distractors that were measured in the pretest. The average time of processing this stage is about 15 minutes.

3.3.4 Comprehension Questions

In order to assess and check the general understanding and some essential details in the two videos, the participants were provided with a comprehension paper test per video. The test contained seven questions, and each question was displayed with three multiple-choice options. Among these questions, the first is about the theme of the video, which is designed to determine whether the subjects are able to completely understand the main idea; the next five questions are all related to the content with detailed information from the clips, while the last question come from one English caption with the target collocation in the video, and asked participants to choose the correct meaning of the target expression according to the context. This process was conducted during the eye-tracking experiment, that is, after watching each video twice, the participants were required to take a set of paper-based comprehension questions.

The researcher scored them based on the two videos. It took about five minutes to finish this test.

3.3.5 Immediate Collocation Posttest

The collocation posttest was designed to evaluate the participants' learning of the target collocations and mainly focused on evaluating how well participants could learn the target collocations after viewing the video twice. Eight multiple-choice questions with three options each was made up of this section, which assessed the participants' ability to understand the meaning of the target collocations. Each sentence provided a context where participants needed to select the most appropriate word from three choices. The multiple choices included both eight target collocations and 16 distractors, for a total of 24 items, allowing researchers to test participants' understanding of the correct collocation. The whole process was completed within an average time of five minutes.

In this study, the pre-test design used a combination of multiple-choice meaning recogonition, fill-in-the-blank questions, and multiple-choice colloation designed to comprehensively assess participants' ability to understand the meanings and uses of the target collocations. This design was intended to establish a more complete baseline for the detection of subsequent learning outcomes. In contrast, the immediate posttest used a multiple-choice question format, with the main purpose of focusing on participants' learning of the target expressions, while improving the consistency and objectivity of scoring by reducing the variables introduced by the task.

Although the differences between the pretest and posttest in terms of question type and measurements may have some impact on the comparability of the test results, the posttest design was aligned with the purpose of the study as the central focus of this study was the learners' improvement in understanding the meanings and uses of the target collocations.

3.4 Apparatus

The eye movements of participants were primarily tracked and monitored by using the Tobii Pro Lab (version 1.61) eye-tracking system, which operates with a 120 Hz sampling rate, ensuring high precision in capturing participants' gaze behaviors. In order for the participants' gaze to naturally align with the area of interest (AOI) on the computer screen, where the stimuli were presented, they were sitting comfortably in a chair that was placed between 60 and 90 cm away from the screen. The screen resolution of the eye tracker's monitor was 1280 x 1024 pixels, which provided clear and high-quality visual stimuli for the participants while ensuring accuracy in tracking even small eye-movements.

During the experiment, participants were instructed to remain as still as possible to minimize head movements, but not required to fix their head or chin on a tabletop or other stationary surface. To ensure the eye movements were tracked accurately, participants were given a brief calibration procedure prior to the experiment, where they followed a series of on-screen targets with their gaze. This helped to synchronize the eye tracker to the individual's eye movements.

3.5 Data Collection

All eye-tracking measurements were collected by Tobii Pro Lab (version 1.61) software, which is a sophisticated eye-tracking system designed for high-precision gaze data acquisition. The focus of this study was narrowed to three key eye-tracking measures—fixation counts, number of visits and total fixation duration—to comprehensively capture learners' allocation of visual attention to target collocations.

A fixation refers to a moment when a participant's gaze remains relatively stationary within a designated Area of Interest (AOI) for at least 100

milliseconds. Fixation counts indicate the total number of such fixation events within a given AOI, which offers insight into the frequency and intensity of learners' focused visual attention. The number of visits is defined as the total number of separate entries into a specific AOI, regardless of whether any fixation occurs during those entries. A single visit may include one or more fixations, or none if the gaze moves through too quickly. By definition, each fixation occurs within a visit, but not every visit contains a fixation. While fixation counts reflect how many times learners focused their gaze within the AOI, number of visits captures how frequently learners returned to the AOI across the entire video, potentially indicating recurring interest or processing attempts. Total fixation duration, on the other hand, is the cumulative time (measured in milliseconds) that a participant's gaze remains within an AOI. Generally speaking, longer fixation duration indicate that an individual is paying closer attention to a specific area, which may be a sign of stronger cognitive processing or trouble comprehending or interpreting the information.

These three measures were selected because they can reveal a more comprehensive picture of how learners allocate their attention, thereby providing valuable and deeper insights into how participants attended to specific visual stimuli during video viewing. However, it is acknowledged that some visits may occur unintentionally (e.g., brief gaze passes during scanning). In this study, all visits were included in the analysis regardless of whether they were intentional or unintentional, as no concurrent verbal reports or recall-based validation methods were employed to distinguish cognitive intent. Therefore, this measure is interpreted with caution. Further discussion of its limitations and implications is provided in the discussion section.

In addition to the eye-tracking data, all participants completed paper-and-pencil tests, which included pretest, comprehension questions and collocation posttest designed to evaluate their learning of the target collocations. These assessments provided a complementary measure for estimating the subjects' collocation learning and served as a standard by which to compare

their performance before, during and after the experiment. The paper-based format was chosen to ensure consistency across participants and to minimize any external variables that might arise from digital testing.

3.6 Statistical Analysis

Statistical analyses for this study were carried out using IBM SPSS Statistics, Version 25 (IBM Corp., Armonk, NY, USA). The primary aim of the analysis was to examine the effect of video viewing order on participants' performance, focusing on how the sequence of caption presentation influenced their comprehension and retention of the enhanced collocations.

The analysis compared two experimental conditions, each representing a different viewing sequence. The participants watched the same videos under different caption settings: one group first viewed the video with English captions featuring enhanced target collocations followed by no captions, while the other group watched the video in the reverse order (i.e., no captions first, then English captions with enhanced target collocations). The viewing order were handled as the main independent variables in this study. The different viewing orders were designed to assess whether the sequence in which captions were presented influenced participants' collocation learning and the understanding of the video content.

The dependent variables in this study included several key measures aimed at evaluating participants' collocation learning and cognitive processing. Pretest scores were used as a baseline measure of participants' prior knowledge of the target collocations, allowing for a comparison of performance between the two groups that experienced different caption presentation orders. Comprehension question scores evaluated participants' comprehension of the video content, giving insight into how well they were able to grasp the meaning and context of the material presented. Collocation posttest scores were analyzed to assess participants' understanding of the target collocations after watching the video,

with an emphasis on their ability to remember and identify the target expressions. These three dependent variables—pretest scores, comprehension question performances, and collocation posttest scores—provided a direct measure of the effectiveness of the video conditions on collocation learning. Meanwhile, other three dependent variables, fixation counts, number of visits and total fixation duration, were collected through eye-tracking data to assess participants' visual attention. It was used to investigate how the visual attention given to the target collocations correlated with performance on the comprehension questions and immediate posttest scores.

3.7 Experimental Procedure

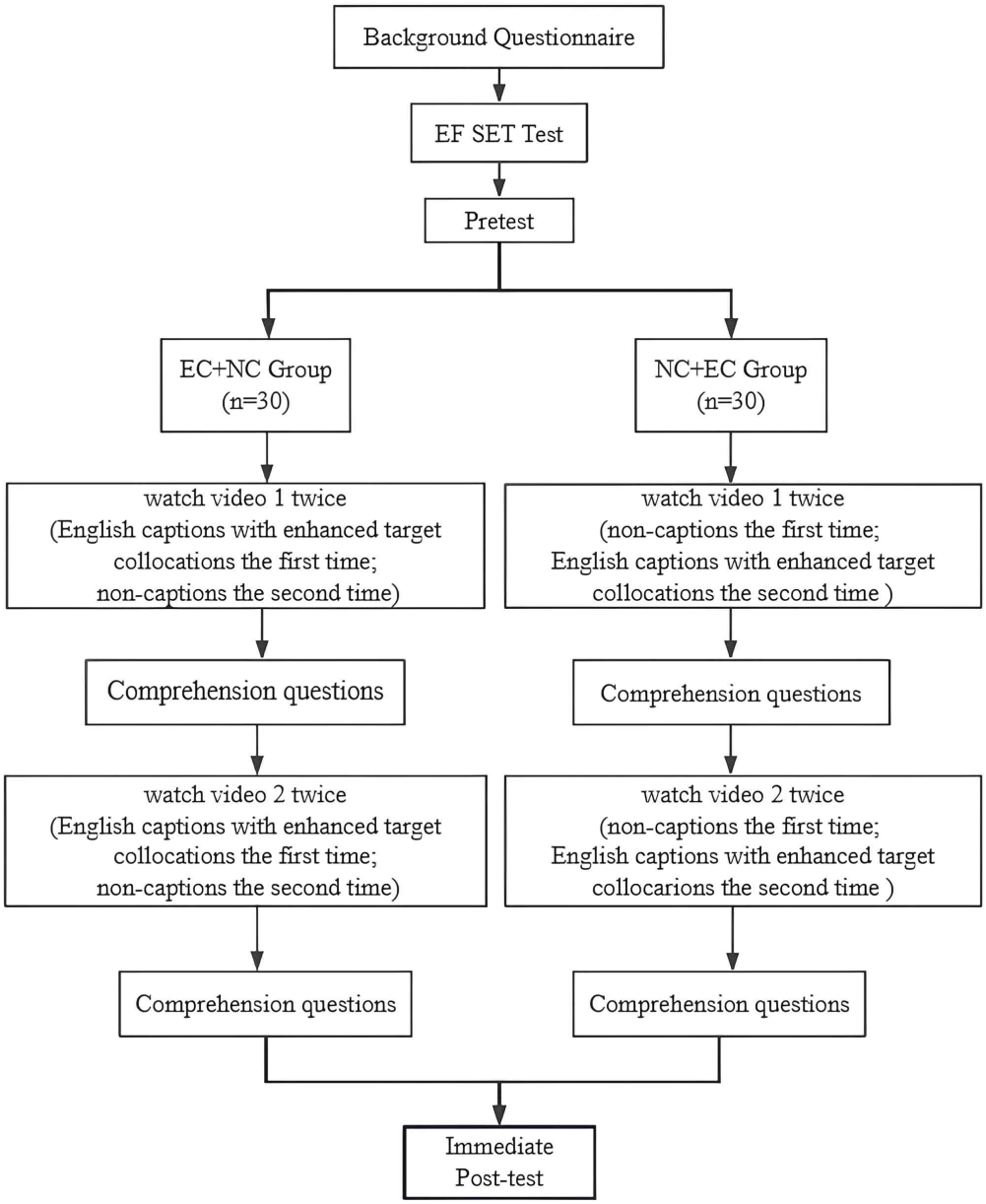

Figure 2 The experimental design

Figure 2 shows the experimental design of this research. Each participant arrived on time in the eye-tracking laboratory within the two-month study period. Upon arrival, they were given a general explanation of the experiment, followed by an opportunity to read and fill out a background questionnaire. This was an essential first step to gather their basic information, such as their gender, age as well as the academic experience in learning English.

Before beginning the experiment, each participant was required to complete an English proficiency test to measure their level of English language skills. The EF SET test included both reading and listening sections designed to assess general English proficiency. After completing the test, participants' scores were converted into levels according to the Common European Framework of Reference for Languages (CEFR).

Then, each participant was informed to take a pretest with three tasks to measure participants' familiarity with the target collocations and their general English ability, providing a baseline for comparison with posttest scores. Then, they were informed that they would be watching two English videos during the experiment, and their eye movements would be tracked and recorded. And they were told that they would be required to answer some questions about the video content afterwards, but they were not given any details about the nature of the content in order to avoid having any effect on their viewing behavior.

Once the pretest was finished, every participant was instructed to sit comfortably in front of the stimulus monitor, positioned at a distance of approximately 60-90 mm from the computer screen. This distance was carefully chosen to ensure optimal calibration of the eye-tracking system. The eye movements of the participants were tracked using the Tobii Pro Lab (version 1.61), equipped with an integrated camera operating at a 120 Hz sampling rate. This system allowed for high-precision recording of the participants' fixation, capturing the smallest eye movements. The screen resolution was set to 1280 x 1024 pixels, ensuring clear visual presentation of the stimuli, while the two videos were displayed in 1280 x 888 pixels to match the aspect ratio of the

content.

Before watching the English videos, each participant underwent a calibration procedure using the Tobii Pro Lab (version 1.61) calibration tool, which allowed the system to accurately track their eye movements. The calibration process involved the participant following a series of on-screen targets with their eyes, and the results were confirmed to guarantee the accuracy of the eye-tracking data. Once calibration was successfully completed, participants were instructed to sit as still as possible to minimize any potential errors and focus on the screen in the tracking process, even though the system allows for some head movements.

The experiment was divided into two different viewing sequence, depending on the participants' assigned group. Sixty individuals were randomly assigned into two groups. One group watched the video with English captions that include enhanced target collocations, followed by the same video without any captions the second time. This viewing order was designed to evaluate how participants engaged with the enhanced target expressions first and how this influence their subsequent attention and learning outcomes when captions were removed. On the contrary, another group began with the English video without captions and then watched the English captioned video with enhanced collocations. This order made it possible to conduct a different comparison, testing whether their first exposure to content without captions affected the collocation learning when enhanced target expressions were later provided.

The two English videos were programmed to start automatically once calibration was completed. After each video began, it played continuously from start to finish without interruption. This ensured that every participant watched the videos in exactly the same way and for the same duration. The videos automatically stopped at the end of each viewing session, marking the completion of that phase of the experiment.

Once the participants finished watching the same video twice under two experimental conditions (English captions with enhanced collocations first or

captions last) each time, they were immediately given a comprehension paper test to assess their general understanding of the video content and some essential details. The researcher scored the tests based on the answers provided for both videos. After the video viewing session, participants were asked to immediately take the collocation posttest with ten multiple-choice questions to measure their understanding of the target collocations. It was designed to evaluate how participants could appropriately use the target collocations after viewing the English videos twice. Following the completion of the experiment, participants were invited to participate in a brief interview to gather feedback on their experiences. The primary focus of the interview was to inquire about the perceived difficulty of the videos and to assess the participants' views on the impact of the caption viewing order (i.e., whether captions containing enhanced collocations were viewed first or last) on their ability to understand the video content.

In order to collect precise eye-tracking data and ensure accurate analysis, the Areas of Interest (AOI) were carefully defined for each type of stimulus presented after the experiment. All AOIs were manually drawn using Tobii Pro Lab software and were rectangular in shape. AOI boxes were drawn around the specific areas of the screen where participants' attention was expected to be focused, namely the English captions and visual images in the videos. Slight variations in AOI size were necessary to accommodate the varying lengths of target expressions. AOIs served as regions for tracking participants' eye movements and collecting the relevant fixation data, which were later used for calculating fixation counts, number of visits and total fixation duration within each area.

For the video that included both captions with enhanced target collocations and visual images, as shown in Figure 3, each caption line and each of the ten target collocations were enclosed within an AOI box respectively that matched the width and height of the caption text, ensuring that the eye-tracking system could accurately record the participant's gaze as it landed on the specific area.

To capture any peripheral gaze activity around the captions and target collocations, an area approximately two letters wider than the text was added as a border around each item. This slight expansion of the AOI box ensured that any fixations close to the edges of the area would still be captured, providing a more comprehensive measure of visual attention to the captioned area. Similarly, AOI boxes were drawn around the visual images, with each box containing the entire area of the image to ensure that all relevant eye movements within the visual regions were accurately tracked.

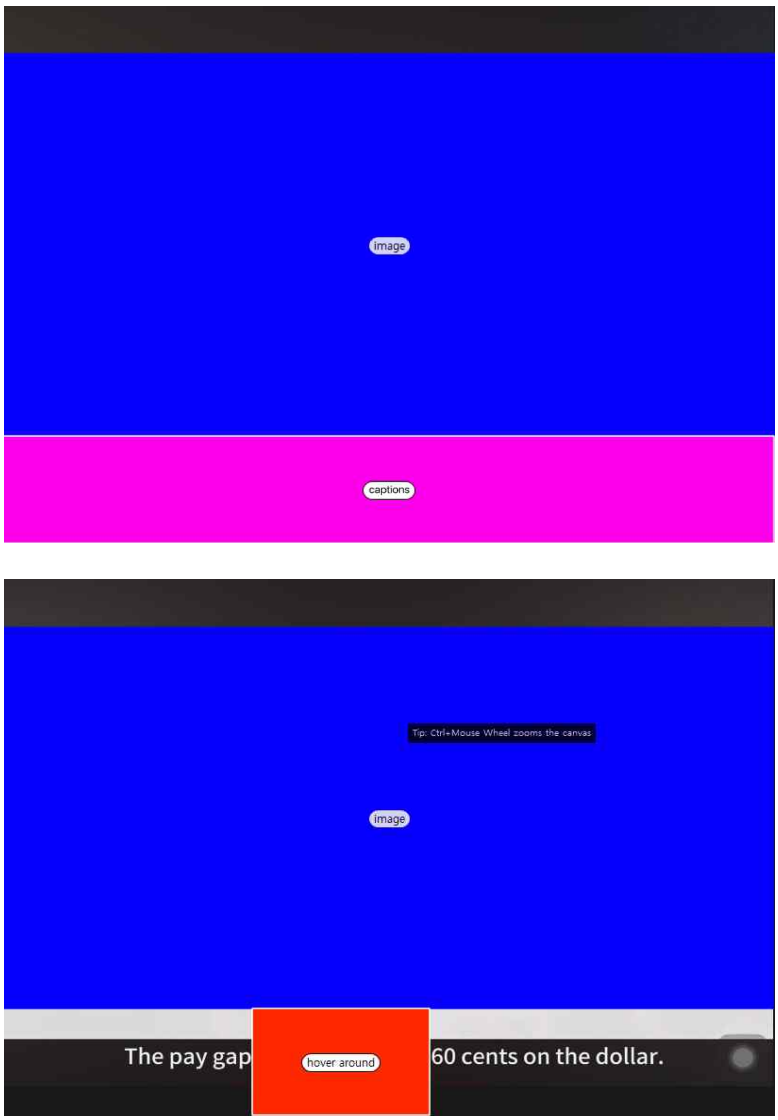

Figure 3 AOI boxes of the video with English captions & target collocations & image

The blue area indicates the image AOI, representing the main visual content of the video. The pink area indicates the caption AOI, corresponding to the location of on-screen captions. The red-highlighted area marks the AOI for one of the enhanced target collocations, *hover around*, which was embedded within the caption. Color coding is used purely for visual distinction and carries no analytical significance. These AOIs were used to extract fixation-related measures including fixation count, number of visits, and total fixation duration.

In the case of the video without any captions, the AOI boxes were drawn only around the visual images (Figure 4). Under this condition, the participants' focus was entirely on the visual content, and the lack of captions meant that only the images were marked as AOI boxes. Similar to the video with English captions, the visual image AOIs were drawn to encompass the entire image. The size of each AOI was adjusted to fit the specific dimensions of the visual element, following the same criteria and adjustment principles as in the caption-present condition.

The format of AOI followed a uniform standard in all experimental conditions, including manual drawing, rectangular shape, size adjustment based on content length, and boundary extension. This standardized approach ensured the transparency, replicablility and comparability of this study in terms of data collection and analysis of results.

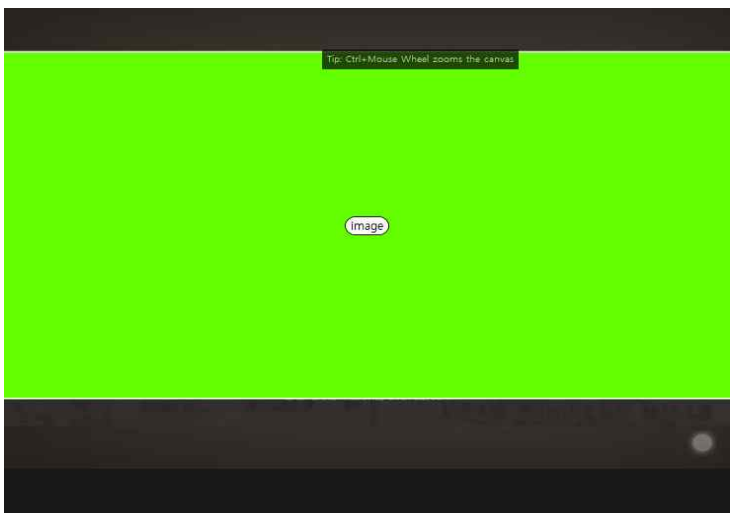

Figure 4 AOI box of the video without captions (image only)

The green area indicates the image AOI used in the non-captioned version of the video. Since no captions were presented in this condition, only the image area was defined for eye-tracking analysis. The color coding is random and used solely for visual distinction.

Since the primary focus of this research is to investigate the effect of caption order on L2 collocation learning, the data related to AOI box of the images in each video will not be analyzed, as the video content itself contains no textual information within visual imagery. Instead, the analysis will concentrate on the AOI boxes of the English captions and the enhanced collocations, regardless of whether captions will be presented in the first or second viewing.

Ethical approval for this study was obtained from the Institutional Review Board (IRB) of Sungshin Women's University (Approval Number: SSWUIRB-2024-035). The study followed ethical guidelines regarding human research and ensured that all participants were informed of the purpose and potential risks of the study, as well as their right to withdraw at any time. All participants provided informed consent before taking part in the experiment.

Chapter 4 Results

4.1 Descriptive Statistics for Pretest, Posttest and Z-scores by Group

This part presents the descriptive statistics for the pre-test, posttest, and standardized (Z) scores of the two experimental groups: EC+NC (English captions with enhanced target collocations first, non-captions second) and NC+EC (non-captions first, English captions with enhanced target collocations second).

As shown in the Tables 5, descriptive statistics indicates that the mean pretest score for Group EC+NC was 38.97 (SD=4.687), while the mean score for Group NC+EC was 39.10 (SD=5.162).

Table 5 Descriptive statistics for pretest scores between the two groups

	Group	N	Mean	Std. Deviation	Std. Error Mean
pretest scores	EC+NC	30	38.97	4.687	.856
	NC+EC	30	39.10	5.162	.942

For posttest scores, Table 6 displays that Group EC+NC had a mean score of 6.70 (SD=1.368). In contrast, Group NC+EC exhibited a higher mean score of 7.63 (SD=0.556). The mean difference between the two groups suggests that learners who watched the non-captioned video first followed by the captioned video performed better on the collocation posttest than those who experienced the reverse sequence.

Table 6 Descriptive statistics for posttest scores between the two groups

	Group	N	Mean	Std. Deviation	Std. Error Mean
posttest scores	EC+NC	30	6.70	1.368	.250
	NC+EC	30	7.63	.556	.102

To enable standardized comparisons of learners' performance across groups, both pretest and posttest scores were converted into Z-scores. As shown in Table 7, for the pretest of Z-scores, the mean values were approximately zero for both groups, with the EC+NC group showing a mean of -0.01 (SD=0.959) and the NC+EC group a mean of 0.01 (SD=1.056). It indicates that participants in the two groups began with comparable levels of baseline proficiency. This further supports the validity of comparing posttest outcomes.

Table 7 Descriptive statistics for standardized pretest and posttest Z-scores between the two groups

	Group	N	Mean	Std. Deviation	Std. Error Mean
Z-scores (pretest)	EC+NC	30	-.01	.959	.175
	NC+EC	30	.01	1.056	.193
Z-scores (posttest)	EC+NC	30	-.41	1.203	.220
	NC+EC	30	.41	.489	.089

In terms of the posttest of Z-scores, the EC+NC group had a mean score of -0.41 (SD=1.203), while the NC+EC group achieved a mean score of 0.41 (SD=0.489). This suggests that, relative to their group averages, participants in the NC+EC group outperformed those in the EC+NC group.

4.2 Statistical Differences in Pretest, Posttest, and Z-scores Between Groups

In order to present the results of inferential statistical analyses conducted to examine group differences in learners' performance on the pretest, posttest, and standardized scores (Z-scores). Independent samples *t*-tests were performed to compare the EC+NC group and the NC+EC group. These analyses aimed to determine whether the order of caption presentation had a statistically significant impact on collocation learning outcomes.

Table 8 *T*-test results for pretest scores between the two groups

		Levene's Test for Equality of Variances		*t*-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
pretest scores	Equal variances assumed	.427	.516	-.105	58	.917	-.133	1.273

The results of the *t*-test (see Table 8) revealed no significant difference in pretest scores between the two groups (t=-0.105, p=.917), with a mean difference of -0.133). These findings suggest that the two groups were equivalent in their baseline proficiency levels, which provides a reliable foundation for subsequent comparisons of posttest performance and eye-tracking measures.

Then, an independent samples *t*-test was conducted to examine whether the sequence in which participants watched English-captioned video with enhanced collocations first or last influenced their posttest performance.

Table 9 *T*-test results for posttest scores between the two groups

		Levene's Test for Equality of Variances		*t*-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
posttest scores	Equal variances not assumed			-3.461	38.323	.001**	-.933	.270

As shown in Table 9, the test revealed a statistically significant difference between the two groups (p=.001) with a mean difference of - 0.933. Since the assumption of equal variances was violated according to Levene's test (p<.001), Welch's ANOVA was conducted and the results are reported in Table 10.

Table 10 Welch's *t*-test results for posttest scores

	Statistic[a]	df1	df2	Sig.
Welch	11.979	1	38.323	.001**

a. Asymptotically F distributed.

The results indicated a significant difference in posttest scores between the EC+NC and NC+EC groups, $F(1, 38.32)=11.979$, $p=.001$.

These two tables indicate that participants in the NC+EC group outperformed those in the EC+NC group, which suggests the non-captioned video first led to greater gains in collocation learning.

To further examine the standardized performance differences, independent samples *t*-tests were conducted on the Z-score transformed pretest and posttest data.

Table 11 *T*-test results for standardized pretest and posttest Z-scores between the two groups

		Levene's Test for Equality of Variances		*t*-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Z-scores (pretest)	Equal variances assumed	.427	.516	-.105	58	.917	-.027	.260
Z-scores (posttest)	Equal variances not assumed			-3.461	38.323	.001**	-.821	.237

As shown in Table 11, for the pretest of Z-scores, the results indicated no significant difference between these two groups ($p=.917$). Levene's test confirmed the assumption of equal variances ($F=0.427$, $p=.516$). These results suggest that the two groups had comparable collocation knowledge prior to the experiment.

In contrast, the posttest of Z-scores revealed a statistically significant difference between groups ($p=.001$). Levene's test indicated unequal variances

(F=24.913, p<.001), and therefore the results were interpreted based on the assumption of unequal variances.

To further validate the group difference using standardized scores, Welch's ANOVA was also performed on the Z-transformed posttest scores. The result is consistent with the Welch's t-test conducted on the raw posttest scores (Table 10), both confirming a significant advantage for the NC+EC group. This finding reinforces the conclusion that the NC+EC condition, where learners first watched the video without captions followed by captions with enhanced target collocations, led to significantly greater gains in collocation learning compared to the EC+NC condition.

To address the first research question, descriptive statistics were also computed for comprehension scores to provide an overview of the data distribution across the two experimental groups (EC+NC and NC+EC). This initial examination aims to determine whether watching the version of captioned video with enhanced collocations first or the non-captioned version first leads to better overall video comprehension.

Table 12 Descriptive statistics for comprehension scores between the two groups

	Group	N	Mean	Std. Deviation	Std. Error Mean
comprehension questions	EC+NC	30	11.03	1.033	.189
	NC+EC	30	12.17	1.206	.220

For comprehension scores, the Table 12 shows that EC+NC group recorded a mean score of 11.03 (SD=1.033). Meanwhile, NC+EC group achieved a higher mean score of 12.17 (SD=1.206). Similar to the posttest scores, this indicates that learners in NC+EC group demonstrated better comprehension of the video content compared to those in EC+NC group.

The results indicate a consistent trend across both measures, with Group NC+EC outperforming Group EC+NC in both posttest scores and comprehension scores.

The two bar charts (Figure 5 & Figure 6) also present the mean posttest

scores and comprehension scores for the two groups, along with their 95% confidence intervals. Both charts consistently show that the NC+EC group outperformed the EC+NC group, with slightly higher mean scores in both measures. This pattern suggests that watching videos without captions first, followed by captions with enhanced collocations, may offer an advantage in improving L2 collocation learning and comprehension.

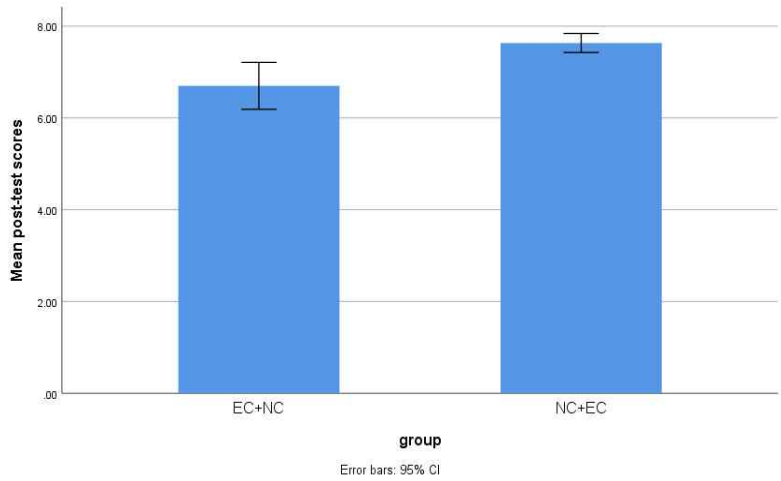

Figure 5 Comparison of mean posttest scores between the two groups

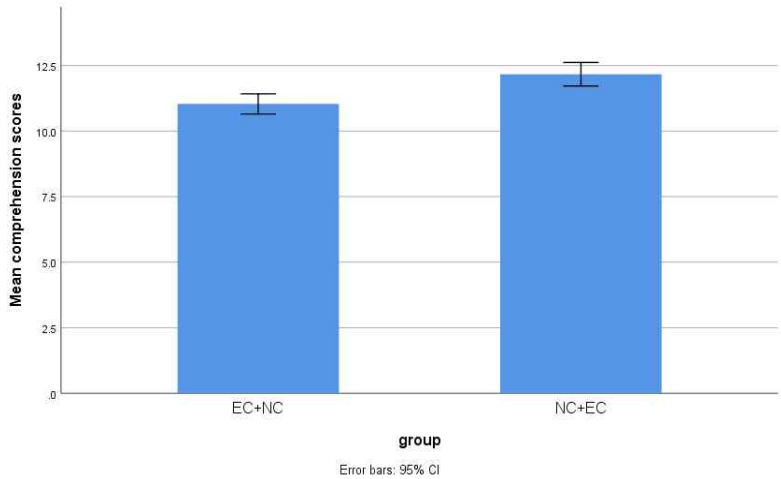

Figure 6 Comparison of mean comprehension scores between the two groups

Table 13 presents the results of an independent samples *t*-test conducted to examine the effects of video viewing sequence on comprehension scores.

Table 13 *T*-test results for comprehension scores between the two groups

		Levene's Test for Equality of Variances		*t*-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
compre-hension scores	Equal variances assumed	1.538	.220	-3.909	58	.000***	-1.133	.290

According to the table, Levene's test indicated that the assumption of equal variances was not violated (p=.220); therefore, results assuming equal variances are reported. The *t*-test revealed a significant difference between the two groups (p<.001). The NC+EC group scored significantly higher in comprehension than the EC+NC group, with a mean difference of -1.133.

The results of the independent samples *t*-tests revealed that the sequence of captions had a significant impact on both their posttest performance and their comprehension of video content. Learners in the NC+EC group—who first watched the video without captions and then with English captions containing enhanced collocations—achieved significantly higher scores than those in the EC+NC group on both the posttest of collocation learning (p=.001) and the comprehension test (p<.001). These findings suggest that delayed exposure to captions with enhanced collocations, following an initial non-captions viewing, promotes deeper processing and facilitates both collocation learning and comprehension.

4.3 Group Differences and Correlations among Eye-Tracking Measures

The second question examines whether learners' attention allocation to enhanced target collocations varies depending on the caption sequence (EC+NC

vs. NC+EC) and explores the interrelationship between different eye-tracking measures. The three key eye-tracking measures considered are fixation counts, number of visits, and total fixation duration, which together provide a comprehensive understanding of learners' visual attention and cognitive processing of the target collocations.

Descriptive statistics for the three eye-tracking measures across the two groups (EC+NC and NC+EC) are presented in Table 14. In terms of fixation counts, the EC+NC group exhibited an average of 560.00 fixation counts (SD=249.470), while the NC+EC group recorded a higher average of 816.90 fixation counts (SD=369.635). For the number of visits to enhanced collocation areas, the EC+NC group had a mean of 13.07 visits (SD=3.532), while the NC+EC group demonstrated a higher mean of 17.20 visits (SD=3.123). For the total fixation duration, the EC+NC group spent an average of 154.351 seconds (SD=117.650) fixating on the enhanced collocation areas, whereas the NC+EC group spent a longer average duration of 268.551 seconds (SD=197.728).

Table 14 Descriptive statistics for three eye-tracking measures in the two groups

	Group	N	Mean	Std. Deviation	Std. Error Mean
Fixation counts	EC+NC	30	560.00	249.470	45.547
	NC+EC	30	816.90	369.635	67.486
Number of visits	EC+NC	30	13.07	3.532	.645
	NC+EC	30	17.20	3.123	.570
Total fixation duraion	EC+NC	30	154.351	117.650	21.480
	NC+EC	30	268.551	197.728	36.100

The results indicate that learners in the NC+EC group demonstrated significantly higher fixation counts, more frequent visits, and longer total fixation duration on the enhanced collocation areas compared to the EC+NC group.

The three boxplots (Figure 7 & Figure 8 & Figure 9) clearly illustrate the distribution of three eye-tracking measures across the two groups.

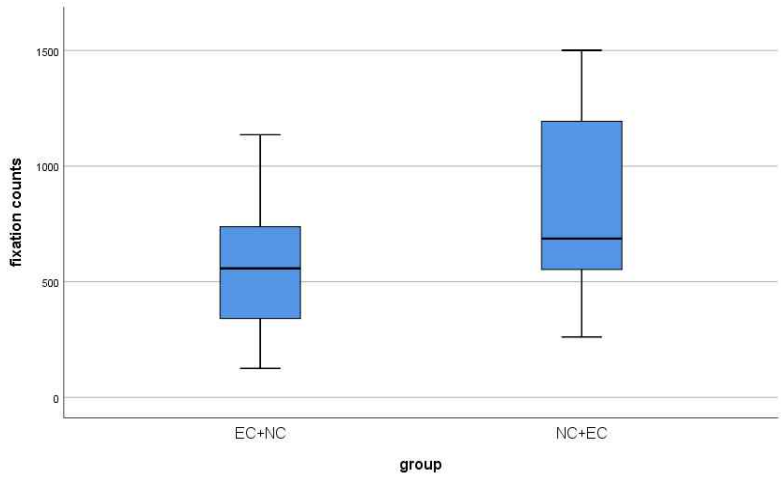

Figure 7 Comparison of fixation counts between the two groups

In terms of fixation counts, Figure 7 shows that participants in the NC+EC group fixated more frequently on the target collocations.

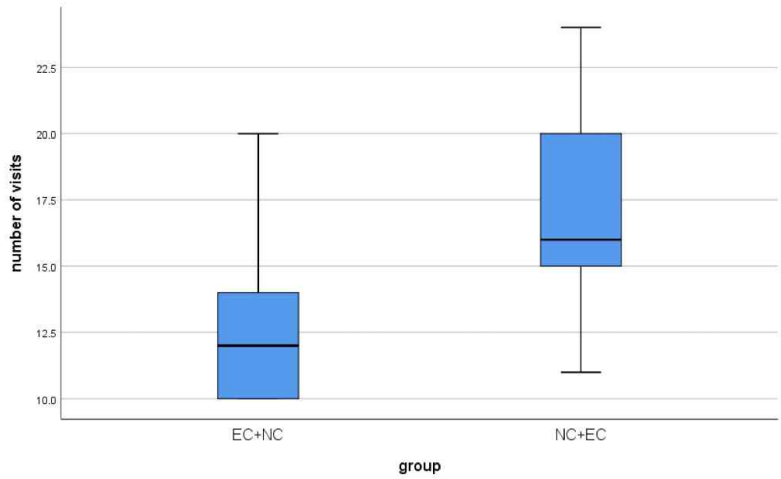

Figure 8 Comparison of number of visits between the two groups

Figure 8 presents the number of visits to the enhanced collocation areas was also higher in the NC+EC group, which suggests that participants were more likely to return to the target areas multiple times for processing.

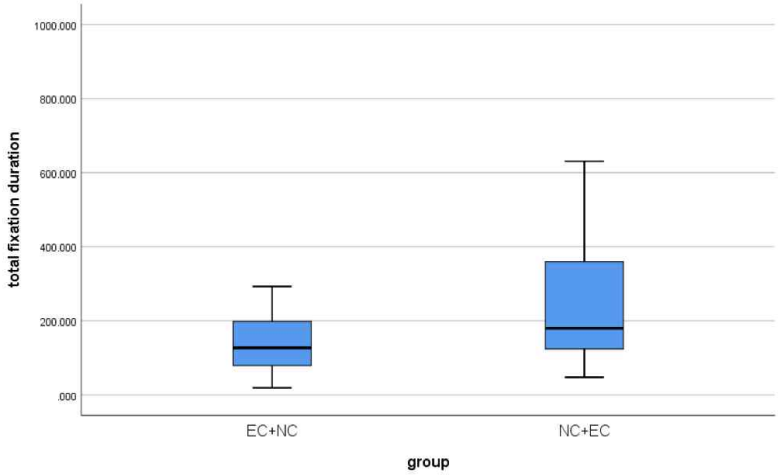

Figure 9 Comparison of total fixation duration between the two groups

The Figure 9 clearly reveals that the total fixation duration was longer for the NC+EC group than for the other group.

Collectively, the figures demonstrate that the NC+EC group tends to show higher engagement in all three measures, while the EC+NC group exhibits more consistency but lower median values. This disparity could suggest that the order of caption presentation influences participant engagement levels, with the NC+EC sequence potentially encouraging more active or varied behaviors.

Further statistical testing was necessary to determine the significance of these differences; therefore, an independent samples *t*-test was conducted to compare the three eye-tracking measures between the two experimental groups (EC+NC and NC+EC). Table 15 shows the results of *t*-test for each measure.

Table 15 *T*-test results for three eye-tracking measures between the two groups

		Levene's Test for Equality of Variances		*t*-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
fixation counts	Equal variances not assumed			-3.155	50.880	.003**	-256.900	81.418
number of visits	Equal variances assumed	.115	.736	-4.802	58	.000***	-4.133	.861
total fixation duration	Equal variances not assumed			-2.719	42.247	.009**	-114.200	42.007

For number of visits, the assumption was met (p=.736), and the equal variance result was used. The result showed a statistically significant group difference (t=-4.802, p<.001). The NC+EC group made more visits to the target areas than the NC+EC group. This suggests that learners in the NC+EC group engaged more frequently with the enhanced collocations compared to the EC+NC group.

Since Levene's test indicated significant violations of the homogeneity of variance assumption for fixation counts and total fixation duration (p<.05), Welch's t-tests were conducted (Table 16).

Table 16 Welch's *t*-test results for fixation counts and total fixation duration

		Statistic[a]	df1	df2	Sig.
fixation counts	Welch	9.956	1	50.880	.003**
total fixation duration	Welch	7.391	1	47.247	.009**

a. Asymptotically F distributed.

Table 16 revealed significant differences between the EC+NC and NC+EC groups in both fixation counts, F(1, 50.880)=9.956, p=.003, and total fixation duration, F(1, 47.247)=7.391, p=.009.

The results indicate that learners in the NC+EC group demonstrated

significantly higher fixation counts, more frequent visits, and longer total fixation duration on the enhanced collocation areas compared to the EC+NC group. These findings suggest that the sequence of video viewing influences learners' attention allocation patterns. That is, learners who watched the non-captioned video first (NC+EC group) appeared to engage more actively with the enhanced collocations, possibly due to increased cognitive effort during the captioned phase after initial exposure to the content.

To further interpret the eye-tracking data, Pearson correlation analysis was used to examine the correlations between the three eye-tracking measures: fixation counts, number of visits, and total fixation duration.

Table 17 Pearson correlations with three eye-tracking measures

		fixation counts	number of visits	total fixation duration
fixation counts	Pearson Correlation	1	.641**	.915**
	Sig. (2-tailed)		.000	.000
	N	60	60	60
number of visits	Pearson Correlation	.641**	1	.504**
	Sig. (2-tailed)	.000		.000
	N	60	60	60
total fixation duration	Pearson Correlation	.915**	.504**	1
	Sig. (2-tailed)	.000	.000	
	N	60	60	60

Note. Correlation is significant at the 0.01 level (2-tailed).

The results (Table 17) reveal significant positive correlations among all three measures, with all coefficients reaching statistical significance at the 0.01 level (two-tailed), suggesting a degree of interdependence in how learners allocate attention to enhanced target collocations during video viewing. The strongest correlation is observed between fixation counts and total fixation duration (r=.915, $p<.001$). This strong positive relationship implies that an increase in the number of fixations on target areas is closely associated with a proportional increase in the total time learners spend fixating on these areas. This finding reflects the consistent relationship between the frequency and cumulative intensity of visual attention allocation. A moderate positive correlation

is found between fixation counts and the number of visits to the target areas (r=.641, p<.001). This relationship indicates that as learners revisit the target areas more frequently, the number of fixations on these areas also increases. While not as strong as the correlation with fixation duration, this finding means that revisits contribute meaningfully to fixation counts. Additionally, another moderate correlation is evident between the number of visits and total fixation duration (r=.504, p<.001), which shows that learners who revisit target areas more often tend to spend slightly more total time fixating on them.

According to these results, the three measures are correlated with each other. It shows a patterns of visual attention allocation when EFL learners are exposed to enhanced target collocations. Although each measure captures a different aspect of attentional engagement, the correlation between them suggests consistency in learners' cognitive processing. Therefore, this correlation analysis was not the focus of this study, but rather as a supplementary analysis to support a more detailed understanding of the mechanisms of learners' attentional allocation during the viewing of captioned videos.

4.4 Correlational and Predictive Relationships Between Eye-tracking Measures and Posttest Performance

The third question further seeks to investigate whether heightened visual attention, as reflected in the eye-tracking measures, and higher comprehension levels directly translate into more successful learning of target collocations. Besides, it aims to determine which combination of these visual variables emerges as the stronger predictor of learners' posttest performance.

Table 18 Pearson correlations between posttest scores and three eye-tracking measures

		fixation counts	number of visits	total fixation duration
posttest scores	Pearson Correlation	.192	.372**	.102
	Sig.(2-tailed)	.141	.003	.439

Note. Correlation is significant at the 0.05 level (2-tailed).

In order to answer this question, Pearson Correlation was run. Table 18 revealed varying degrees of association. Among the measures examined, only the number of visits displayed a statistically significant correlation with posttest scores (r=.372, p=.003), revealing a positive relationship. This indicates that learners who revisited the enhanced target collocations more often tended to achieve better performance in the posttest. On the other hand, neither fixation counts nor total fixation duration demonstrated any significant correlation with posttest scores, which means that the frequency and total time spent visually attending to the target areas were not closely linked to learning outcomes. These results emphasize that revisiting target areas might facilitate enhanced cognitive processing or repeated exposure to key collocations. Consequently, it can be inferred that how learners allocate their visual attention—such as revisiting target collocations—may play a more critical role in successful collocation learning.

The multiple regression analysis was employed to examine the extent to which three eye-tracking measures—fixation counts, number of visits, and total fixation duration—predict learners' posttest performance on enhanced target collocations.

Table 19 Model summary for multiple regression analysis predicting posttest scores from three eye-tracking measures

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.391[a]	.153	.107	1.075

Note. Predictors: (Constant), fixation counts, number of visits, total fixation duration

As reported in the model summary (Table 19), the regression explained approximately 15.3% of the variance in posttest scores (R^2=.153), with an adjusted R^2 of .107. The standard error of the estimate was 1.075, suggesting that, on average, predicted scores deviated from the actual posttest scores by just over one point on the 10-point scale.

Table 20 ANOVA[a] for multiple regression analysis predicting posttest scores from three eye-tracking measures

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11.647	3	3.882	3.361	.025*
	Residual	64.687	56	1.155		
	Total	76.333	59			

Note. Dependent Variable: posttest scores
Predictors: (Constant), fixation counts, number of visits, total fixation duration

The ANOVA result (Table 20) further confirms the statistical significance of the model (F=3.361, p=.025), and indicates that the model accounted for a significant amount of variance in learning outcomes.

Table 21 Coefficients[a] for multiple regression analysis predicting posttest scores from three eye-tracking measures

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.406	.574		9.427	.000***
	fixation counts	.001	.001	.181	.510	.612
	number of visits	.113	.048	.387	2.331	.023**
	total fixation duration	-.002	.002	-.259	-.821	.415

Note. Dependent Variable: posttest scores

An examination of the standardized regression coefficients (see Table 21) revealed that only number of visits emerged as a statistically significant predictor of posttest performance (β=.387, t=2.331, p=.023). This suggests that learners who revisited the enhanced target collocation areas more frequently tended to perform better in the collocation posttest, even after controlling for other forms of visual attention. In contrast, neither fixation counts (β=.181, p=.612) nor total fixation duration (β=−.259, p=.415) were significant predictors when all three variables were entered into the model, which indicates that the frequency or duration of visual attention alone did not explain additional variance in learning outcomes.

Furthermore, the effect size for the regression model, calculated using Cohen's f^2, was approximately 0.181, which reveals a medium effect (Cohen, 2013). This suggests that the model accounts for a meaningful proportion of the variance in posttest performance.

These findings were further supported by the partial regression plot, which visually demonstrated the unique contribution of each predictor. Among the three eye-tracking measures, only the partial plot for number of visit revealed a clear positive linear trend, confirming its predictive value (Figure 10).

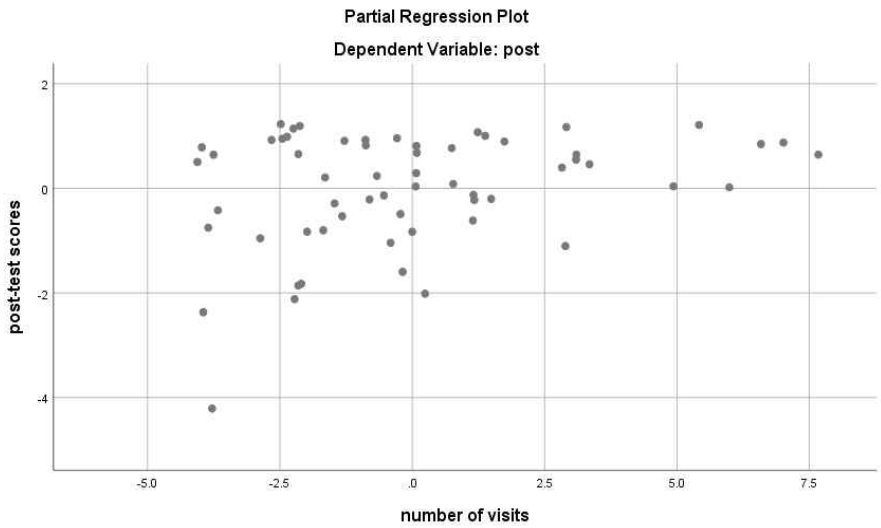

Figure 10 Partial regression plot for the number of visits predicting posttest scores

The other two predictors—fixation counts and total fixation duration—showed no meaningful linear pattern when controlling for the influence of the remaining variables. It indicated that the frequency and total time of attention allocation to target areas did not significantly contribute to learning outcomes.

The findings show that the number of visits may be a more meaningful indicator of learning processes than fixation counts or total fixation duration. This could be attributed to the cognitive reprocessing facilitated by revisiting target areas multiple times. While fixation counts and total fixation duration reflect the overall intensity of attention allocation, they do not necessarily capture the strategic nature of learners' engagement with the content. Repeated visits may indicate a deeper level of cognitive engagement, as learners likely re-encounter the collocations in varied contexts or reinforce their memory through repetition. Additionally, the lack of significance for fixation counts and total fixation duration may suggest that merely looking at target areas, even for extended periods, is insufficient for successful learning unless it is coupled with meaningful processing and interaction with the content.

Chapter 5 Discussion

5.1 Research Question 1

The results of this question provide compelling evidence that the sequence in which EFL learners watched English videos (captions with enhanced collocations first vs. non-captions first) significantly influences their L2 collocation learning and comprehension. The findings revealed that the participants in the NC+EC group consistently outperformed their counterparts in the EC+NC group on both posttest scores and comprehension scores.

The descriptive and inferential statistical analyses demonstrated that the NC+EC group achieved higher posttest scores and exhibited greater comprehension scores, compared to the EC+NC group. The independent sample *t*-test results further confirmed the significant difference of the video viewing sequence on posttest and comprehension performance. The NC+EC sequence appears to be more effective in facilitating collocation learning, possibly because learners in this condition were able to consolidate their understanding of target collocations, while watching the captioned video with enhanced collocations phase, having first been exposed to the content without captions. Additionally, this sequence seems to enhance overall comprehension, as learners likely engaged more deeply with the video content in the second phase, building on their initial experience with the non-captioned version. In contrast, participants in the EC+NC group, who were first exposed to the version of the captioned video with enhanced target collocations, may have become overly reliant on the captions. This reliance could have reduced their cognitive engagement with the broader visual and auditory content of the video. As a result, the cognitive processing required to integrate visual, contextual, and linguistic information may have been less effective in this group, leading to lower overall engagement and, ultimately, poorer learning outcomes. The NC+EC sequence allows learners to

initially focus on contextual and visual cues, and foster deeper cognitive processing as they actively infer meaning. The subsequent exposure to English captions with enhanced collocations then reinforces their understanding, not only of the target collocations but also of the overall video content, by providing explicit linguistic input that builds on their prior contextual knowledge.

The results are partially consistent with prior research highlighting the advantages of non-captioned videos in fostering deeper processing. For instance, Peters and Webb (2018) noted that learners were able to acquire vocabulary when watching videos in the non-captions condition, especially for frequent and visually salient words. However, they did not directly explore the effect of the order in which captions were presented, nor did they compare the different conditions with and without captions. Therefore, this study builds on their findings by further examining the role of caption presentation order in collocation learning. Also, Winke, Gass, and Sydorenko (2010) found that caption presentation order can influence learning outcomes, but the results varied according to the learners' language background. This suggests that although this study found some advantages of NC+EC order, the optimal effect of caption order may depend on individual learner characteristics rather than being universally applicable. Similarly, Vanderplank (2016) highlighted that captions can serve as a dual-modality aid, but their efficacy depends on learners' prior familiarity with the content and language. By watching non-captioned videos first, learners may have developed a foundational understanding of the video content, which subsequently facilitated the deeper processing of collocations when enhanced captions were introduced. Therefore, while the current study partially aligns with previous research demonstrating the potential for incidental learning and the relevance of caption order, it extends this line of inquiry by systematically examining the sequencing effect on L2 collocation learning using enhanced captions.

5.2 Research Question 2

The second research question offers strong evidence that, in comparison to the EC+NC sequence, the NC+EC sequence leads to significantly higher attention allocation to enhanced target collocations. By analyzing three key eye-tracking measures—fixation counts, number of visits, and total fixation duration—the results of this question highlights the importance of sequencing captioned and non-captioned videos to optimize learners' cognitive engagement with target collocations.

The significant differences in fixation counts, number of visits, and total fixation duration between the two groups suggest that the NC+EC sequence promotes more active cognitive processing of target collocations. Participants in this group, having first watched the non-captioned video, were likely compelled to rely on contextual and visual cues to infer meaning, which might have primed them to focus more intensively on enhanced collocations when captions were introduced. The findings resonate with the dual-route hypothesis proposed by Holsanova et al. (2009), which argues that visual and verbal processing pathways interact dynamically during multi-modal learning. The NC+EC group capitalizes on this interaction by first encouraging learners to rely on visual and contextual cues, thereby activating their inferential processing mechanisms. When captions with enhanced collocations are introduced in the second stage, the linguistic input reinforces previously inferred meanings, leading to heightened attention allocation to enhanced collocations. This dual-phase engagement strategy may explain why learners in the NC+EC group demonstrated greater cognitive focus on target collocations.

The results align with findings from Winke et al. (2013), who demonstrated that multimodal input encourages learners to process language both visually and cognitively. The NC+EC sequence appears to optimize this process by supporting learners' attention from a less enriched phase (non-captioned) to a more input-enhanced one (captioned with enhancements). In

contrast, the EC+NC group may have relied too heavily on captions during the first phase, which could have reduced their need to process visual and contextual cues actively. This reduced cognitive effort in the initial phase may have contributed to lower engagement with enhanced collocations in the subsequent non-captioned phase, as reflected in their lower fixation measures.

Furthermore, it also extends the findings of Bisson et al. (2012), who showed that captions can direct learners' visual attention to specific linguistic features, thereby enhancing retention. However, the present study demonstrates that the sequence in which captions are introduced significantly modulates this effect. By first exposing learners to non-captioned videos, the NC+EC group reduces the likelihood of over-reliance on textual input, which is a common drawback of initial captioned exposure. In contrast, learners in the EC+NC group may have experienced a reduced need to actively process visual and contextual information during the initial captioned phase, which leads to lower engagement with enhanced collocations in the subsequent non-captioned phase.

It is worth noticing that the positive correlations among the three eye-tracking measures suggest a degree of interdependence in how learners allocate attention. The strong correlation between fixation counts and total fixation duration ($r=.915$, $p<.001$) implies that learners who fixate more frequently on target areas also tend to spend more cumulative time on these target areas. This finding aligns with previous studies (e.g., Godfroid et al., 2013) that highlight the relationship between fixation duration and cognitive engagement. Furthermore, the moderate correlations between number of visits and other two eye-tracking measures also indicate that revisits to target areas contribute significantly to overall fixation activity, though not all revisits result in proportionally higher fixation counts and longer fixation duration. These relationships reinforce the idea that attention allocation is a multidimensional process influenced by both frequency and intensity of engagement.

5.3 Research Question 3

The analysis of the third research question highlights the relationship between visual attention measures and learning outcomes for target collocations. Among the three eye-tracking measures analyzed, only number of visits to enhanced collocation regions showed a significant positive correlation with posttest scores (r=.372, p=.003). This result was further supported by multiple regression analysis, which showed that the combination of fixation counts, number of visits, and total fixation duration explained 15.3% of the variance in posttest performance (R^2=.153). Among the three predictors, only number of visits emerged as a significant predictor (p=.023), which reveals that learners who revisited the enhanced collocation areas more frequently tended to achieve better learning outcomes. In contrast, fixation counts and total fixation duration were not significant predictors, which suggests that simply allocating visual attention—whether in terms of frequency or duration—is insufficient for effective learning without repeated and purposeful cognitive engagement.

This finding seems unexpected, as fixation counts are generally considered to be a better indicator of sustained attention than number of visits. However, there are several possible explanations for this result. First of all, fixation counts reflect localized, focused attention during a single visit, but may not capture the broader, repetitive cognitive engagement that occurs over the entire time period. In contrast, the number of visits represent learners' tendency to re-expose themselves to the same language item in different contexts or at different stages of comprehension. Even if some visits are unintentional or incidental, the cumulative effect of multiple exposures may provide additional opportunities for retrieval, or contextual elaboration-processes that are critical for long-term retention. Secondly, it is also possible that fixations, which is the indicative of visual attention, do not consistently translate into effective learning due to the high cognitive load associated with multimodal input. Learners need to interpret visuals and captions while processing auditory input. As a result, sustained

fixations may occur in the absence of meaningful processing, and repeated visits may reflect learners' attempts to refine or reassess their understanding in the briefly available time.

The significant positive relationship between the number of visits and posttest scores aligns with the retrieval practice theory (Roediger & Butler, 2011), which emphasizes the importance of repeated exposure and retrieval opportunities for memory consolidation. Revisiting target areas likely provided learners with multiple instances to rehearse and solidify their understanding of collocations, enhancing long-term retention. This finding complements previous research by Fisher and Frey (2014), who found that re-visitation could strengthen learners' ability to process contextualized language. Similarly, the elaborative encoding theory (Craik & Tulving, 1975) suggests that deeper processing of input enhances memory retention. Repeated visits to target collocations may encourage learners to engage in elaborative processing, as they encounter collocations in varied linguistic and contextual contexts. This finding extends the work of Hattie and Timperley (2007), who emphasized the role of feedback loops in reinforcing learning. Re-visitation may act as a self-regulated feedback mechanism, and enable learners to evaluate and refine their understanding of collocations.

Contrary to earlier studies emphasizing the role of total fixation duration in vocabulary learning (e.g., Holmqvist et al., 2011), this section found no significant relationship between total fixation duration and immediate posttest performance. One possible explanation is that in multimedia learning contexts, learners often need to divide their attention across multiple modalities, including visual, auditory, and textual inputs. This division of cognitive resources may dilute the potential impact of sustained fixations on specific visual targets. While learners may direct their visual attention to enhanced collocations, this allocation does not necessarily guarantee deeper cognitive processing (Godfroid, Boers, & Housen, 2013). As a result, the effect of total fixation duration on learning outcomes may be diminished, which emphasizes the distinction between passive

attention and active, meaningful engagement with the content. This observation is consistent with the findings of Kormos (2012), who addressed the potential limitations of passive attention in complex language tasks. And it also aligns with Moreno and Mayer's (2007) cognitive theory of multimedia learning, which highlights the challenges of managing cognitive resources in multi-modal environments.

Fixation counts, similarly, were not predictive of posttest performance, suggesting that frequency alone does not guarantee meaningful engagement. This result is consistent with studies by Dörnyei and Skehan (2003), which argue that attention alone is insufficient to support effective learning unless it is accompanied by intentional learning strategies such as conscious rehearsal, or noticing of form-meaning relationships. In contrast, the number of visits to target areas emerged as a significant predictor of posttest performance. While not all visits may represent intentional processing—some may result from unintentional entries into the AOI—their cumulative frequency likely reflects increased opportunities for attention, reprocessing, and repeated exposure. In this sense, the number of visits can be interpreted as an indirect indicator of learner engagement, especially in multimodal or dynamic learning environments.

Overall, these findings stress the need to reconsider the role of attention measures in predicting language learning outcomes. While eye-tracking has been widely used to measure attention allocation (e.g., Godfroid et al., 2013), this research question demonstrates that not all attention measures are equally meaningful. While fixation counts and durations capture intensity within a single gaze episode, the number of visits may better reflect dynamic processing behavior across time.

However, it is important to note that the number of visits analyzed in this study included all recordings of access to the area of interest (AOI), regardless of whether the participants intended to do so or not. In other words, visits may have occurred intentionally—learners' conscious return to the target expression or unintentionally—simply be the result of unintentional eye movements during

the viewing of the video. Therefore, although the result shows that an increased number of visits may mean that learners have more opportunities to notice and engage with the target expressions, it cannot be explicitly assumed that every visit corresponds to intentional cognitive processing. Future research may benefit from applying minimum fixation duration time within each visit or from integrating eye-tracking data with other cognitive monitoring technologies, in order to better distinguish between intentional attention and incidental attention.

Chapter 6 Conclusion

6.1 Major Findings

The main purpose of this study was to investigate how the order of English caption presentation (captions with enhanced collocations first vs. last) affects collocation learning in L2 learners. By systematically varying the two sets of video conditions for two groups of participants, this study utilized eye-tracking technology with the aim of determining whether presenting captions with enhanced collocations earlier or later in the video affected participants' eye-movement data as well as their ability to comprehend the videos and retain the target collocations.

The three central research questions in this study were designed to explore the relationship between caption order, learners' attention, comprehension, and collocation learning, each building on the others in a hierarchical manner. The first question is *To what extent does the sequence in which EFL learners watch English-captioned and non-captioned videos (watch English-captioned videos with enhanced collocations first or later) influence their immediate posttest performance on enhanced target collocations and their video comprehension scores?* This foundational question sets the stage for understanding whether caption presentation order affects learning outcomes. Building on this, the second question—*To what extent does the order of captioned and non-captioned video viewing result in distinct patterns in learners' eye-tracking measures (fixation counts, number of visits, and total fixation duration) to enhanced target collocation areas and to what extent are these three eye-tracking measures interrelated?*—focused on the learners' attentional allocation in response to caption sequences and whether these measures were correlated with each other, thus providing insight into how learners engage with multimodal input. The third question *To what extent are*

eye-tracking measures (fixation counts, number of visits, and total fixation duration) correlated with immediate posttest scores, and to what extent do they collectively predict EFL learners' final performance on enhanced target collocations? links these findings by exploring the correlations between eye-tracking measures and immediate posttest scores, and assesses how these attention-related measures work together to predict a learner's final performance on the target collocations. Together, these three questions form a hierarchy that first identifies the effect of caption order and group differences on learning outcomes, then examines the role of attention allocation as reflected in eye-tracking data, and finally investigates the predictive relationships between attention, comprehension, and final performance.

The findings revealed that several key insights regarding the effects of caption order on both learners' attention and learning outcomes. First of all, the order in which the videos were viewed had a significant main effect on both immediate posttest scores in collocation and video comprehension scores. That is, viewers in the NC+EC group who watched the non-captioned video first followed by captions with enhanced collocations consistently outperformed and more effective in facilitating target collocation learning than viewers in the EC+NC group (who watched the captions containing enhanced collocations first) in terms of posttest performance and comprehension scores. Second, the NC+EC sequence led to significantly greater attention allocation to the target collocations compared to the EC+NC sequence. Significant differences between the two groups on three eye-tracking measures showed that the NC+EC sequence encourages more active cognitive processing of the target collocation. And also, positive correlations between the three eye-tracking measures suggest a degree of interdependence in the allocation of learners' attention. Third, among the eye-tracking measures analyzed, the number of visits to the target collocation area was the only one that showed a significant positive correlation with the results of the immediate posttest in both groups. This finding suggests that participants who visited the enhanced colloations more frequently during video

viewing performed better on the posttest, which highlights the potential role of repeated visual engagement in supporting the collocation learning.

6.2 Implications

6.2.1 Theoretical Implications

The findings of this study make an important contribution to the theoretical understanding of second language (L2) learning and multimodal learning, particularly in the areas of collocation learning and caption processing. This study extends existing theories by exploring how the sequencing of English captions (captions with enhanced collocations first vs. last) affects collocation learning performance, a topic that has received relatively little attention in the literature.

The sequence of introducing visual and textual cues only after the learner has been exposed to the context through auditory input is consistent with the progressive scaffolding approach, an instructional methodology that progressively increases the level of support provided to the learner. In this approach, learners first encounter auditory input without captions, attempting to understand the content without visual aids, thereby activating their prior knowledge and contextual inference abilities, enhancing cognitive engagement, and activating deeper processing mechanisms. Then, captions with enhanced contextual cues are introduced to reinforce and consolidate prior understanding. This approach contrasts with traditional approaches, which tend to provide captions from the outset, which may distract learners from the auditory and visual contexts. Therefore, this study questions the assumption that captions as a standalone resource can maximize L2 learning outcomes and suggests that the order of caption exposure is a key factor in optimizing collocation learning.

In addition, the eye-tracking method used in this study provides new insights into the relationship between attention and collocation retention. The

positive correlation between the number of visits to the enhanced collocation area and posttest scores may lend indirect support to the noticing hypothesis (Schmidt, 1990), which emphasizes the role of conscious awareness in language learning. While not all visits can be assumed to reflect intentional or conscious attention, the repeated visit to specific collocations likely increased the opportunity for noticing to occur. Besides, the results of the present study add to the growing body of research on the role of multimodal input in L2 learning, supporting the view that visual cues combined with auditory input can enhance the learning process, especially for more complex forms like collocations (VanPatten, 2004).

Lastly, this study contributes to the understanding of collocation learning, which is an underexplored area compared to other aspects of vocabulary acquisition. By focusing on collocations, the study emphasizes the importance of contextualized learning for second language learners, and suggests that exposure to target expressions in rich multimedia contexts can improve their understanding of collocations and their functional use in communication. These findings highlight that successful collocation learning involves not only recognizing individual words, but also understanding how they co-occur and interact meaningfully in specific contexts.

6.2.2 Pedagogical Implications

The results of this study have important implications for the design of video-based English language learning materials. Firstly, the findings suggest that sequencing video materials from non-captioned to captioned English versions optimizes learners' engagement and retention of the target language forms. This order allows learners to rely on visual and auditory context first, followed by textual support to solidify their comprehension. The introduction of enhanced captions after the initial exposure to auditory and visual contexts reinforces collocation learning. This sequence progressively deepens language

processing. In addition, the study highlights the importance of eye-tracking data for assessing learners' attention allocation, which can help educators to adjust instructional strategies to better support learners' interaction with multimedia materials. By analyzing measures such as fixation counts, number of visits, and total fixation duration, educators can gain insights into the depth and focus of learners' visual attention to specific language forms and enable them to adapt the design of video materials to emphasize key linguistic elements. Moreover, the findings underscore the importance of revisiting in learning, and imply that interactive features—user-controlled elements embedded in digital learning materials that allow learners to actively engage with the content—can encourage learners to revisit and reinforce key target language forms, particularly collocations. Example of such features include clickable definitions (enable learners to click on a word or phrase to instantly view its meaning, usage, or translation), and replayable clips (allow learners to replay specific video segments). Finally, the research suggests that learners can benefit from guided support and metacognitive training on how to engage effectively with enhanced input, ensuring that visual enhancement is used purposefully to maximize retention of collocations.

6.3 Limitations and Future Directions

Although the present study provides strong evidence for the efficacy of the NC+EC sequence in facilitating collocation learning and video comprehension, it is important to acknowledge some of its limitations.

Firstly, the sample size, while sufficient for statistical analyses, may limit the generalization of the findings to larger or more diverse L2 populations. A larger, more heterogeneous sample would help to confirm whether the observed effects are consistent across learners.

Secondly, this study used short video clips to present the learning material, which may not fully replicate the complexity or richness of a real

viewing experience, where language is exposed in longer, or more varied contexts. Future research could investigate whether the effects of caption order persist across longer videos or more challenging materials, which could provide a more realistic understanding of how caption order affects learning in real-world environments.

Thirdly, the selected collocations may differ slightly in terms of frequency and collocation type. Despite the measures taken to control for them in this study, subtle differences in phonological features and learners' past exposure experience may still have an impact on attention allocation and learning outcomes to some extent. Future research could more systematically control for differences in perceptual salience and familiarity of target collocations and such variables to better isolate the effect of caprions order.

Fourthly, while the use of eye-tracking technology has provided valuable insights into learners' attention allocation, these data could be further analyzed to explore more detailed patterns of attention and how they relate to caption processing. Future studies may look at a wider range of eye-tracking measures such as first fixation duration or second fixation duration on specific linguistic features, to gain more insight into how attention is allocated across multimodal inputs during learning. Morever, it is important to note that one of the key eye-tracking measures used in this study—the number of visits to the target collocations—did not distinguish between intentional and unintentional gaze entries. Since the number of visits recorded was based only on gaze entry into the AOIs, it may include both purposeful re-engagement and unintentional fixations. This limits the ability to draw definitive conclusions about learners' conscious processing of target collocations. Future research may consider integrating minimum fixation thresholds, or retrospective reporting techniques to better distinguish between intentional attention and unintentional visual contact. Furthermore, although the eye-tracker with 120 Hz sampling rate that used in this study are suitable for video viewing tasks, its accuracy may not be sufficient to capture rapid eye movements with high precision (Conklin,

Pellicer-Sánchez, & Carrol, 2018). Future research could use higher frequency devices and technologies with 500 Hz or 1000 Hz to better distinguish between intentional and unintentional attention.

Fifthly, the present study focused only on collocation, while other important aspects of language learning, such as grammar, pronunciation or pragmatic competence, were not investigated. Therefore, future research could expand the scope of the investigation to explore whether similar sequencing effects can be observed when learning other language objectives or in different instructional contexts, and it may provide a more comprehensive understanding of the benefits of caption sequencing. Longitudinal studies would also be valuable to examine the long-term retention of collocations obtained through different caption sequencing conditions, which would shed light on the persistence of learning outcomes over time.

Another limitation is the difference between the pretest and posttest in terms of test format and measurement, which may limit the direct comparability of the results to some extent. Future research may consider using a consistent test format in the pretest and posttest to enhance data comparability.

Finally, individual differences, such as motivation, cognitive style, and language proficiency levels may mediate the effects observed in this study. Investigating these factors could lead to a more comprehensive understanding of how learners allocate attention and process multimodal input. By examining how cognitive and ability-related variables affect the learning process, future research can refine instructional strategies to better meet the needs of diverse learners and increase the effectiveness of caption-based interventions.

REFERENCES

Abbasian, G. R., & Yekani, N. (2014). The role of textual vs. compound input enhancement in developing grammar ability. *Issues in Language Teaching*, *3*(1), 134-113.

Alotaibi, H. M., Mahdi, H. S., & Alwathnani, D. (2023). Effectiveness of subtitles in L2 classrooms: A meta-analysis study. *Education Sciences*, *13*(3), 274.

Alqaed, M. A. (2022). *The effect of an explicit awareness-raising approach on lexical collocation awareness and knowledge* (Doctoral dissertation, University of Leicester).

Askari, H. (2024). To teach or not to teach collocations in EFL academic contexts: An overview of current research and a response to Reynolds (2019, 2022). *The Reading Matrix: An International Online Journal*, *24*(2), 62-113.

Benson, M., Benson, E., & Ilson, R. (1997). *The BBI dictionary of English word combinations*(Rev. ed.). Amsterdam & Philadelphia: John Benjamins Publishing Company.

Bisson, M. J., Van Heuven, W. J., Conklin, K., & Tunney, R. J. (2012). Processing of native and foreign language subtitles in films: An eye tracking study. *Applied Psycholinguistics, 35*(2), 399-418.

Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. (2014). Gauging the effects of exercises on verb - noun collocations. *Language Teaching Research*, *18*(1), 54-74.

Choi, S. (2017). Processing and learning of enhanced English collocations: An eye movement study. *Language Teaching Research*, *21*(3), 403-426.

Choi, S. (2023). Visual saliency in captioned digital videos and learning of English collocations: An eye-tracking study. *Language Learning & Technology*, *27*(1), 1-21.

Cintrón-Valentín, M. C., & García-Amaya, L. (2021). Investigating textual

enhancement and captions in L2 grammar and vocabulary: An experimental study. *Studies in Second Language Acquisition*, *43*(5), 1068-1093.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Routledge.

Conklin, K., & Pellicer-Sánchez, A. (2016). Using eye-tracking in applied linguistics and second language research. *Second Language Research, 32*(3), 453-467.

Conklin, K., Pellicer-Sánchez, A., & Carrol, G. (2018). *Eye-tracking*. Cambridge: Cambridge University Press.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268-294.

Danan, M. (2004). Captioning and subtitling: Undervalued language learning strategies. *Meta*, *49*(1), 67-77.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193-210.

Dörnyei, Z., & Skehan, P. (2003). Individual differences in second language learning. *The handbook of second language acquisition*, *1*(18), 589-630.

Fazlali, B., & Shahini, A. (2019). The effect of input enhancement and consciousness-raising techniques on the acquisition of lexical and grammatical collocation of Iranian EFL learners. *TESL-EJ*, *24*(2), 1-21.

Fisher, D., & Frey, N. (2014). Content area vocabulary learning. *The Reading Teacher*, *67*(8), 594-599.

Firth, J. R. (1957). *Papers in linguistics 1934 - 51*. Oxford: Oxford University Press.

Gass, S. M. (1997). *Input, interaction, and the second language learner*. New York: Routledge.

Godfroid, A. (2020). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. New York: Routledge.

Godfroid, A., Ahn, J., Choi, I. N. A., Ballard, L., Cui, Y., Johnston, S., Lee, S., Sarkar, A., & Yoon, H. J. (2018). Incidental vocabulary learning in a natural reading context: An eye-tracking study. *Bilingualism: Language and Cognition*, *21*(3), 563-584.

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye-tracking. *Studies in Second Language Acquisition*, *35*(3), 483-517.

Goldschneider, J. M., & DeKeyser, R. M. (2001). Explaining the "natural order of L2 morpheme acquisition" in English: A meta analysis of multiple determinants. *Language learning*, *51*(1), 1-50.

Goudarzi, Z., & Moini, M. R. (2012). The effect of input enhancement of collocations in reading on collocation learning and retention of EFL learners. *International Education Studies*, *5*(3), 247-258.

Han, Z., Park, E. S., & Combs, C. (2008). Textual enhancement of input: Issues and possibilities. *Applied Linguistics*, *29*(4), 597-618.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Hoey, M. (2012). *Lexical priming: A new theory of words and language.* New York: Routledge.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford: Oxford University Press.

Holsanova, J., Holmberg, N., & Holmqvist, K. (2009). Reading information graphics: The role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *23*(9), 1215-1226.

Huang, H. C., & Eskey, D. E. (1999). The effects of closed-captioned television on the listening comprehension of intermediate English as a second language (ESL) students. *Journal of Educational Technology Systems*, *28*(1), 75-96.

Huang, Q., Abdul Samat, N., & Haladin, N. A. B. (2024). The role of exposure condition, awareness and item type in developing implicit and explicit knowledge of collocational rules. *Cognitive Processing*, *25*(3), 403-420.

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language learning*, *51*(3), 539-558.

Jung, J. & Lee, M. (2024). Incidental collocational learning from reading-while-listening and the impact of synchronized textual enhancement. *International Review of Applied Linguistics in Language Teaching*, *62*(4), 1935-1958.

Jung, J., Stainer, M. J., & Tran, M. H. (2022). The impact of textual enhancement and frequency manipulation on incidental learning of collocations from reading. *Language Teaching Research.* Advance online publication. https://doi.org/10.1177/13621688221129994.

Kho, S.Q.E., Aryadoust, V. & Foo, S. (2023). An eye-tracking investigation of the keyword-matching strategy in listening assessment. *Educ Inf Technol* 28, 3739-3763.

Kim H., Choi S., Kweon, S-O. (2023). Effects of announcing a vocabulary test before reading a glossed text on reading behaviors and vocabulary acquisition: An eye-tracking study. *PLOS ONE 18*(1): 1-18.

Kormos, J. (2012). The role of individual differences in L2 writing. *Journal of second language writing*, *21*(4), 390-403.

Krashen, S. (1985). *The input hypothesis: Issues and implications*. London: Longman Press.

Krishnamurthy, R. (2006). Collocations. In *Encyclopedia of language and linguistics. 7*(2), 596-600.

Laufer, B., & Waldman, T. (2011). Verb noun collocations in second language writing: A corpus analysis of learners' English. *Language learning*, *61*(2), 647-672.

Lee, M., & Jung, J. (2024). Effects of textual enhancement and task manipulation on L2 learners' attentional processes and grammatical knowledge

development: A mixed methods study. *Language Teaching Research*, *28*(4), 1552-1571.

Lee, M., & Révész, A. (2020). Promoting grammatical development through captions and textual enhancement in multimodal input-based tasks. *Studies in Second Language Acquisition*, *42*(3), 625-651.

Lee, S. K., & Huang, H. T. (2008). Visual input enhancement and grammar learning: A meta-analytic review. *Studies in second language acquisition*, *30*(3), 307-331.

Li, H., Paterson, K. B., Warrington, K. L., & Wang, X. (2022). Insights into the processing of collocations during l2 English reading: evidence from eye movements. *Frontiers in Psychology*, *13*, 845590.

Li, L. X. (2023). Promoting accuracy of collocation use in L2 writing: the role of data-driven learning in indirect corrective feedback. *Computer Assisted Language Learning*, Advance online publication. https://doi.org/10.1080/09588221.2023.2292554.

Majuddin, E., Siyanova-Chanturia, A., & Boers, F. (2021). Incidental acquisition of multiword expressions through audiovisual materials: The role of repetition and typographic enhancement. *Studies in Second Language Acquisition*, *43*(5), 985-1008.

Markham, P. L., Peter, L. A., & McCarthy, T. J. (2001). The effects of native language vs. target language captions on foreign language students' DVD video comprehension. *Foreign Language Annals*, *34*(5), 439-445.

Mayer, R. E. (2014). Incorporating motivation into multimedia learning. *Learning and Instruction*, *29*, 171-173.

Mertzen, D., Paape, D., Dillon, B., Engbert, R., Vasishth, S., & Vasishth, S. (2023). Syntactic and semantic interference in sentence comprehension: Support from English and German eye-tracking data. *Glossa Psycholinguistics*, *2*(1), 1-48.

Michael, L. (1993). *The Lexical approach: The state of ELT and a way forward*. Hove: Language Teaching Publications.

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of abnormal psychology*, *110*(1), 40-48.

Montero Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014). Effects of captioning on video comprehension and incidental vocabulary learning. *Language Learning & Technology*, *18*(1), 118-141.

Montero Perez, M., Peters, E., & Desmet, P. (2018). Vocabulary learning through viewing video: the effect of two enhancement techniques. *Computer assisted language learning*, *31*(1-2), 1-26.

Montero Perez, M., Van Den Noortgate, W., & Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System,* 41(3), 720-739.

Nation, I. (2006). How large a vocabulary is needed for reading and listening?. *Canadian Modern Language Review*, *63*(1), 59-82.

Nation, I. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Neuman, S. B., & Koskinen, P. (1992). Captioned television as comprehensible input: Effects of incidental word learning from context for language minority students. *Reading Research Quarterly*, 95-106.

Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford: Oxford University Press.

Park, H., Choi, S., & Lee, M. (2012). Visual input enhancement, attention, grammar learning, & reading comprehension: An eye movement sudy. *English Teaching*, *67*(4), 241-265.

Peters, E. (2009). Learning collocations through attention-drawing techniques: A qualitative and quantitative analysis. *Researching collocations in another language: Multiple interpretations*, 194-207.

Peters, E. (2012). Learning German formulaic sequences: The effect of two attention-drawing techniques. *The Language Learning Journal*, *40*(1), 65-79.

Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition*, *40*(3), 551-577.

Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition from and while reading: An eye-tracking study. *Studies in Second Language Acquisition*, *38*(1), 97-130.

Pellicer-Sánchez, A., Siyanova-Chanturia, A., & Parente, F. (2022). The effect of frequency of exposure on the processing and learning of collocations: A comparison of first and second language readers' eye movements. *Applied Psycholinguistics*, *43*(3), 727-756.

Perez, M. M., Van Den Noortgate, W., & Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System*, *41*(3), 720-739.

Puimège, E., Montero Perez, M., & Peters, E. (2023). Promoting L2 acquisition of multiword units through textually enhanced audiovisual input: An eye-tracking study. *Second Language Research*, *39*(2), 471-492.

Puimège, E., Montero Perez, M., & Peters, E. (2024). The effects of typographic enhancement on L2 collocation processing and learning from reading: An eye-tracking study, *Applied Linguistics, 45*(1), 88-110.

Rajendran, K., & Mustafa, H. R. B. (2023). Vocabulary acquisition through graded captioned videos among lower proficiency second language learners. *International Journal of Research and Innovation in Social Science*, *7*(8), 1269-1286.

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*(8), 1457-1506.

Robinson, P. (2003). Attention and memory during SLA. *The handbook of second language acquisition*, 631-678.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences*, *15*(1), 20-27.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching1. *Language Teaching*, *47*(4), 484-503.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129-158.

Simard, D. (2009). Differential effects of textual enhancement formats on intake. *System*, *37*(1), 124-135.

Sinclair, J. (1987). "Collocation: A progress report", in Steele and Threadgold (eds.), 319-331.

Sinclair, J. (1987). The dictionary of the future. *Library Review*, *36*(4), 268-278.

Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: Acquisition of collocations under different input conditions. *Language Learning*, *63*(1), 121-159.

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied linguistics*, *16*(3), 371-391.

Szudarski, P., & Carter, R. (2016). The role of input flood and input enhancement in EFL learners' acquisition of collocations. *International Journal of Applied Linguistics*, *26*(2), 245-265.

Teng, M. F. (2020). *Language learning through captioned videos: Incidental vocabulary acquisition*. New York: Routledge.

Teng, M. F., & Cui, Y. (2025). Second language collocation learning through captioned videos: how do learners' vocabulary knowledge and working memory affect learning?. *Computer Assisted Language Learning*, 1-29.

Tomlin, R. S., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in second language acquisition*, *16*(2), 183-203.

Toomer, M., & Elgort, I. (2019). The development of implicit and explicit knowledge of collocations: A conceptual replication and extension of Sonbul and Schmitt (2013). *Language Learning*, *69*(2), 405-439.

Vanderplank, R. (2016). 'Effects of' and 'effects with' captions: How exactly does watching a TV programme with same-language subtitles make a difference to language learners? *Language Teaching*, *49*(2), 235-250.

VanPatten, B. (2004). Input processing in second language acquisition. In VanPatten, B. (Ed.), *Processing instruction: Theory, research, and commentary* (pp.5-31). Mahwah, NJ: Lawrence Erlbaum Associates.

Vilkaitė, L., & Schmitt, N. (2019). Reading collocations in an L2: Do collocation processing benefits extend to non-adjacent collocations? *Applied Linguistics*, *40*(2), 329-354.

Vu, D. V., & Peters, E. (2022). Incidental learning of collocations from meaningful input: A longitudinal study into three reading modes and factors that affect learning. *Studies in Second Language Acquisition*, *44*(3), 685-707.

Vu, D. V., & Peters, E. (2023). A longitudinal study on the effect of mode of reading on incidental collocation learning and predictors of learning gains. *TESOL Quarterly*, *57*(1), 5-32.

Wang, Y. (2019). Effects of L1/L2 captioned TV programs on students' vocabulary learning and comprehension. *Calico Journal*, *36*(3), 184-203.

Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology*, *14*(1), 65-86.

Winke, P., Gass, S., & Sydorenko, T. (2013). Factors influencing the use of captions by foreign language learners: An eye-tracking study. *The Modern Language Journal*, *97*(1), 254-275.

Xie, H., Mayer, R. E., Wang, F., & Zhou, Z. (2019). Coordinating visual and auditory cueing in multimedia learning. *Journal of Educational Psychology*, *111*(2), 235.

Zagar, D., Pynte, J., & Rativeau, S. (1997). Evidence for early closure attachment on first pass reading times in French. *The Quarterly Journal of Experimental Psychology Section A*, *50*(2), 421-438.

Zhang, G., Yuan, B., Hua, H., Lou, Y., Lin, N., & Li, X. (2021). Individual differences in first-pass fixation duration in reading are related to resting-state functional connectivity. *Brain and Language*, *213*, 104893. https://doi.org/10.1016/j.bandl.2020.104893.

APPENDIX

1. Video Materials

Video 1: *Why women are paid less*

When I was growing up,
I knew one woman lawyer. One.
I never met a woman doctor.
I couldn't have even imagined women engineers.
The pay gap hovered around 60 cents on the dollar.
It was caused by several interconnected factors,
like lower female education rates,
women not being in the workforce in big numbers,
grouping in traditionally feminine industries,
and the fact that it was perfectly legal to pay women less,
and then a slew of cultural norms about gender roles and aptitudes.
These were the major explanations for the pay gap.
And then, in just a few decades, things changed.
Sisterhood is powerful! Join us now!
The battle cry of the women's liberation movement
rings out down New York's Fifth Avenue.
First woman to receive the highest honor of the National...
The House broke into spontaneous applause.
Benazir Bhutto, the new prime minister.
This is the first American woman in space.
The first woman ever nominated to the Supreme Court.
The first woman ever to run on a Presidential ticket.
My candidacy has said to women, "The doors of opportunity are open."
Women are out-earning men in college degrees and advanced degrees.
Women are being engaged to bring the next generation.

Maybe for the first time in history,
women are actually outnumbering men in the workplace.
This was just a sea change
to see women competing for scholarships I couldn't have competed for,
going to schools that were not open to women,
taking on jobs that were close to women.
That's changed...just...unbelievably.
Many of the factors that were causing the pay gap shrunk,
except for one.
But what has stayed is that women bare children.
They are assumed to be the primary caregiver.
Even as women became doctors, and lawyers, and heads of state,
the popular expectation remained in society
that they would still do most of the work of raising children.
In the United States, in the UK, even in progressive Scandinavian countries,
surveys today show that only a fraction of the population
thinks women should work full-time when they have young kids.
When it comes to men, the expectation flips.
70 percent of Americans think that new fathers should work full-time.
There still is a considerable percentage of people,
not just in our country, but around the world,
who really think once you're a mom, you shouldn't be in the workplace.
And that's been proven wrong, short-sighted over and over again.
I learned, after I went back, when my time was constrained,
not by my employer, but by me,
because I wanted to get home to that baby
and spend time with her,
that I could actually get a lot of work done in 15 minutes.
Like, I would take any opportunity to work.
I've become, I think, a much better employee since I've had children.

But even when a mother does work full-time, just like her male partner,
she spends nine hours a week more than him on childcare and housework.
Over a year, that's the equivalent of an extra three months of a full time job.
This is the heart of the pay gap.
And to understand why, it helps to follow the story
of a young couple just starting out on their careers.
I often think about
the trajectories of the many law students I taught.
They look exactly the same.
They have the same educational record, the same experience.
And then you watch what starts to happen
as they hit their late 20s, early 30s, childbearing years,
and then they start thinking about having children.
If they have children, at that point, somebody has to be home.
You can have lots of child care,
but you know a parent needs to be at home for those situations that needs a parent.
So he's likely to get promoted.
She, on the other hand, has had to turn down some of those assignments,
say no to some of that travel.
So eight years out, ten years out,
typically, he's then a partner, and he can do lots of things from there.
She hasn't made partner. She's not earning the same.
She's working flexibly, or even part-time,
and from there, her earning potential and his just keep diverging.
This is the story the data tells us in study after study
in a variety of different countries.
One Danish study did an especially good job
of showing how childbirth affects earnings.
Here's a man's pay trajectory.

Watch what happens when his child is born.
Here's the woman's trajectory.
So then if you compare the earnings of a woman with kids
to a woman without kids,
you can see that the pay gap isn't as much about being a woman
as it is about being a mom.
The gender gap really is between women with children and everybody else.

Video 2: *World's water crisis*

NASA satellite data shows aquifers in northern India
decreasing by 29 trillion gallons in just a decade.
There are simply more people on Earth consuming more water.
This century, water consumption has increased seven-fold.
And the rain and snow that we count on to water crops and refill lakes and rivers
is getting less reliable.
Climate change is making available water much more erratic.
We're seeing areas around the world
that are experiencing much more extended dry periods.
But the problem isn't just that there's more people on Earth using water,
it's how we're using water.
Humans need to drink almost a gallon of water per day.
Brushing your teeth, washing your hands, typically uses about a gallon.
There goes three gallons.
But the drinking, washing and toilet flushing
of every person on Earth only accounts for 8% of our freshwater use each year.
Most of the water goes to agriculture and industry,
and into the food and products we use.
Let's take a bottle of Coca Cola.
90% of the water in that bottle
is not what you see in that bottle.
98% of the water is actually embedded in all the ingredients that were grown to make that bottle of Coca Cola.
74 liters of water goes into every glass of beer.
A cup of coffee 130 liters.
Each of your cotton shirts 2,500 liters.
But nothing has as much embedded water as meat.
Alfalfa is a common ingredient in cattle feed,

and growing a kilogram of it takes 510 liters of water.
An average cow consumes about 12 kilograms of feed a day.
Divided up,
just one quarter pound hamburger takes around 1,650 liters of water to produce.
The world is eating more and more like Americans.
Higher calorie diets with more meat.
But everyone can't eat like Americans.
There actually isn't enough water in the world.
Water doesn't abide by some of the basic rules of capitalism.
Farmers hardly pay anything for it.
So the true cost of water doesn't end up in the cost of the burger.
Which is why those fast food places can offer you bargain burgers.
How can it be 99 cents?
For only 2.99. You heard right. 2.99.
In most places in the world,
water is treated and priced like there will always be enough of it.
So we end up using it in absurdly wasteful ways.
Arid Southern California uses over 2 trillion gallons of water a year
to grow alfalfa, which they get from the Colorado River,
hundreds of miles away.
The amount they pay for it doesn't even cover the cost of delivery.
Just a fraction of the water used by South Africa's wine industry
would be enough for Cape Town's taps.
India and China both grow their most water-intensive crops
in some of their driest regions.
But as water gets more scarce, that may change.
The bank Goldman Sachs predicted that water would be
the petroleum of the 21st century.
And private interests, like hedge funds, have started buying up water,
prompting fears that they'll take advantage of scarcity to turn a profit.

And if that sounds like a villain's plot in a James Bond movie,
that's because it was.
As of this moment,
my organization owns more than 60% of Bolivia's water supply.
This contract states that your new government
will use us as utilities provider.
But putting a higher price on water might have benefits.
The benefit of valuing water as we should
and sending, you know, a price signal,
is that we wouldn't be growing alfalfa in the desert.
Remember that point. It'll be important later.
We wouldn't be growing crops that don't make sense in really arid places.
Because the economics of it wouldn't make sense.
And 95% of the irrigated farmland in the world
probably wouldn't use the most inefficient irrigation method...
just flooding the fields.
And if water had a higher price,
governments might decide it's worth the money
to repair our water infrastructure.

2. Background Questionnaire

Thank you for participating in this experiment. This questionnaire is used to collect some background information of participants as part of this study. All information will be kept confidential and only be used for academic purposes.

1. Name: _____
2. Gender: _____
3. Age: _____
4. Home country: _____
5. Native language: _____
6. Academic major: _____
7. TOEIC/TOEFL/IELTS/CET-4/CET-6 score: _____
8. Have you lived in an English-speaking country for more than one month in the past? If yes, please provide details such as the country, the duration of stay and so on.

9. At what age did you first learn English through formal education or private tutoring?

10. Do you watch English videos with English captions or not? How often do you watch English videos with or without English captions?

11. Have you participated in any eye-related programs? If yes, please provide details.

3. Pretest

Name : _____

1. Choose the option (A, B or C) that best represents the meaning of each vocabulary collocation from the list.

1.keep pace with	A. Move quickly. B. Move at the same rate as. C. Move slowly.
2.the trajectories of	A. The specific tools used in a particular profession. B. The individual elements that make up a whole. C. The progressions or paths of something over time.
3.sense of security	A. Feeling of safety. B. Feeling of fear. C. Feeling of uncertainty.
4.by chance	A. Deliberately. B. Coincidentally. C. Accidentally.
5.take into account	A. Consider. B. Disregard. C. Ignore.
6.alarms ring	A. Alarms make a sound. B. Alarms cause fear. C. Alarms stop ringing.
7.a bunch of	A. A specific group of. B. A small amount of. C. A large number of.
8.make an impression	A. Create a physical mark. B. Create a positive or lasting effect. C. Fail to be noticed.

9.strong tea	A. Flavored tea. B. Weak tea. C. Intensely flavored tea.
10.keep diverging	A. Maintain a constant rate. B. Move further apart over time. C. Stay close together.
11.deeply impressed	A. Not affected emotionally. B. Slightly moved emotionally. C. Extremely moved emotionally.
12.common ingredient	A. A frequently used or typical component in a recipe. B. An unusual and rare component in cooking. C. A decorative element added for presentation.
13.reflect on	A. Glare at. B. Mirror. C. Consider deeply.
14.balanced diet	A. Equal amounts of food. B. Varied and nutritious food. C. Limited food choices.
15.reject an appeal	A. Accept an appeal. B. Decline an appeal. C. File an appeal.
16.heavy rain	A. Intense rain. B. Light rain. C. Brief rain.
17.catch a bus	A. Miss a bus. B. Drive a bus. C. Board a bus.

18.highly unlikely	A. Very improbable. B. Very probable. C. Possibly likely.
19.depend on	A. Rely on. B. Dislike. C. Ignore.
20.apologize profusely	A. Apologize insincerely. B. Apologize excessively. C. Apologize reluctantly.

2.Fill in each blank with the most appropriate collocation from the word bank provided.

break the ice	lifelong learning	water consumption	keep in mind	a pile of
a big deal	a slew of	get prepared	beautiful scenery	save time
primary caregiver	social media	fall asleep	get more scarce	take a nap

1. The new environmental regulations aim to reduce _____ in urban areas.
2. The concept of _____ is essential in today's rapidly changing world.
3. Winning that award was _____ for her career in journalism.
4. Using a planner can help you _____ and stay organized throughout the day.
5. She has been the _____ since her mother passed away last year.
6. To avoid last-minute stress, you should _____ for your presentation well in advance.
7. It's important to _____ that deadlines are strict in this project.
8. The _____ from the top of the mountain was absolutely breathtaking.
9. As resources _____, the need for sustainable practices becomes more urgent.

10. The report highlighted _____ reasons why the project failed to meet its objectives.

11. There was _____ laundry on the bed that needed to be folded.

12. Many people use _____ to stay connected with friends and family.

13. During the meeting, the manager told a funny story to _____ and make everyone feel comfortable.

14. After reading a book for a while, I usually _____ very quickly.

15. After working all morning, I decided to _____ to recharge for the afternoon.

3. Choose the correct collocation (A, B or C) to complete the sentences.

1. He gained _____ the secret files.

   A. access to

   B. access about

   C. access with

2. She has a strong _____ music.

   A. interest about

   B. interest for

   C. interest in

3. The temperature in the room tends to _____ 22 degrees Celsius.

   A. hover with

   B. hover around

   C. hover about

4. The author's philosophy is _____ every chapter of the book.

   A. embedded in

   B. embedded by

   C. embedded for

5. There is growing _____ climate change.

   A. concern of

   B. concern with

   C. concern about

6. The _____ his absence is unknown.

A. reason with

B. reason to

C. reason for

7. It is crucial to _____ the safety regulations.

A. abide by

B. abide about

C. abide in

8. Lucy had an _____ his sister.

A. argument about

B. argument in

C. argument with

9. I listened to a _____ child poverty.

A. debate to

B. debate at

C. debate about

10. We are all in favour of a _____ fireworks.

A. ban for

B. ban on

C. ban with

11. One _____ this method is its simplicity.

A. advantage for

B. advantage of

C. advantage to

12. We hope that one day they will find a _____ this horrid disease.

A. cure for

B. cure of

C. cure about

13. Her _____ the research was invaluable.

A. contribution of

B. contribution with

C. contribution to

14. The _____ success is hard work.

A. key about

B. key to

C. key of

15. He has a _____ classical music.

A. preference for

B. preference of

C. preference in

4. Comprehension Questions

Name : _____

Video 1: *Why Women Are Paid Less?*

1. What is the main topic of this video?

A. The history of women's rights movements.

B. The impact of women in leadership positions.

C. The reasons for the gender pay gap.

2. What are some factors that contributed to the pay gap according to the video?

A. Higher female education rates.

B. Women being in the workforce in big numbers.

C. Lower female education rates and cultural norms about gender roles.

3. What significant changes occurred due to the women's liberation movement?

A. Women continued to earn less in education.

B. Women began out-earning men in college degrees and advanced degrees.

C. Women remained in traditionally feminine industries.

4. What expectation remained in society even as women became doctors, lawyers, and heads of state?

A. Women should still do most of the work of raising children.

B. Women should be the primary breadwinners.

C. Women should not work at all.

5. What did the Danish study mentioned in the video reveal about the impact of childbirth on earnings?

A. Childbirth increases earnings for women.

B. Parenthood affects earnings trajectories differently for men and women.

C. Parenthood has no impact on earnings.

6. What is the main reason for the pay gap between men and women?

A. Being a mother.

B. Being a woman.

C. Being a wife.

7. In the video, the phrase “a slew of cultural norms” is used. What does “a slew of” mean in this context?

A. A specific type of.

B. A small number of.

C. A large quantity of.

Video 2: *Water Crisis*

1. What is the main topic of this video?

A. The benefits of drinking water.

B. The world's water crisis.

C. The history of water usage.

2. Which factor makes available water much more erratic, according to the video?

A. Increased industrialization.

B. Growing population.

C. Climate change.

3. How much water does human daily drinking and washing account for in terms of freshwater use each year?

A. 8%.

B. 15%.

C. 25%.

4. Which product mentioned in the video has the highest amount of embedded water?

A. Meat.

B. Coffee.

C. Cotton shirts.

5. Why is water usage in agriculture and industry a significant concern?

A. It is the most regulated.

B. It costs the most money.

C. It uses a large portion of the available freshwater.

6. What might be a positive outcome of putting a higher price on water, according to the video?

A. Increased industrial profits.

B. More extensive water bottling.

C. Less wasteful use of water.

7. In the video, there is one sentence "Water doesn't abide by some of the

basic rules of capitalism." What does "abide by" mean in this context?

A. To benefit something for profit.

B. To follow or obey a rule or principle.

C. To criticize or challenge a system openly.

5. Posttest

Name : _____

1. Choose the best answer (A, B or C) to complete the sentences below.

1. The artist's work reflects _____ of her personal experiences and emotional growth over the years.

A. the trajectories of

B. the principles of

C. the rules of

2. Sugar is a(an) _____ in many processed foods, contributing to health concerns due to its high consumption.

A. rare material

B. ordinary product

C. common ingredient

3. The report highlighted _____ new discoveries in the field of medicine, showcasing a range of breakthroughs.

A. a few

B. a slew of

C. a handful of

4. _____ in urban areas has been a growing concern, leading to initiatives to promote water conservation.

A. Air pollution

B. Water consumption

C. Environmental destruction

5. In many families, the mother is often the _____, taking care of the children and managing household duties.

A. primary caregiver

B. chief officer

C. main leader

6. Citizens are expected to _____ the laws of their country to maintain order and justice.

A. break down

B. hold off

C. abide by

7. Prices for agricultural products often _____ certain levels due to market fluctuations and supply chain disruptions.

A. wander around

B. circle around

C. hover around

8. Memories of his childhood home are _____ his mind, bringing him comfort in times of stress.

A. separated from

B. embedded in

C. located at

6. Data

Data of EC+NC group

participants	pretest	comprehension questions	post-test	fixation counts	number of visits	total fixation duration
1	24	10	6	393	10	81.325
2	35	9	5	541	12	127.307
3	35	12	3	584	11	126.157
4	39	11	8	901	20	292.757
5	43	12	7	188	11	24.135
6	43	10	8	488	21	88.984
7	41	11	6	270	10	48.919
8	44	12	6	312	10	68.881
9	45	10	8	968	21	255.599
10	44	12	8	437	11	79.085
11	40	11	7	271	10	53.156
12	38	11	6	1136	20	536.074
13	34	11	7	805	13	206.041
14	37	11	8	668	12	134.094
15	35	12	5	699	12	151.620
16	45	12	8	619	11	148.048
17	33	10	5	336	10	85.931
18	40	11	8	574	13	158.841
19	42	11	6	919	15	479.455
20	40	9	8	341	10	78.301
21	35	12	7	662	16	207.361
22	40	9	5	125	10	19.246
23	37	11	8	356	10	105.664
24	37	11	5	528	14	153.542

25	35	11	5	324	10	67.432
26	38	12	8	757	14	198.314
27	40	12	8	825	13	194.366
28	39	10	7	587	14	127.229
29	45	12	7	738	17	233.468
30	46	13	8	448	11	99.195

Data of NC+EC group

participants	pretest	comprehension questions	post-test	fixation counts	number of visits	total fixation duration
1	47	13	8	911	21	237.948
2	40	14	8	461	11	87.951
3	40	13	8	1249	21	325.631
4	48	11	8	553	21	123.975
5	27	13	7	428	18	103.208
6	30	11	8	709	14	256.178
7	41	12	8	559	16	144.001
8	38	13	7	1052	17	335.736
9	43	10	8	785	24	213.822
10	31	11	7	402	19	79.219
11	40	13	7	1501	22	613.072
12	40	12	8	727	16	185.867
13	40	12	8	1346	22	407.944
14	46	14	8	1465	15	868.758
15	35	13	6	1245	17	630.625
16	37	10	7	1194	16	359.345
17	33	14	7	1266	19	494.904
18	40	11	8	332	15	82.090
19	43	13	8	1030	21	347.977
20	36	12	7	591	14	155.503
21	46	13	8	1017	16	475.391
22	36	11	8	629	14	167.707
23	34	11	8	628	15	162.667
24	38	10	7	662	14	152.720

25	43	12	8	664	16	173.560
26	33	12	8	393	14	90.788
27	43	13	7	583	16	120.988
28	40	12	8	261	17	47.583
29	41	14	8	1350	20	461.437
30	44	12	8	514	15	149.933

ABSTRACT

In second language (L2) learning, the increasing use of multimedia resources has sparked growing interest in the role of captions in enhancing collocation learning. However, despite many studies focusing on the general benefits of captions for language learning, little is known about the impact of the order in which captions are presented. This research explored the impact of the order of enhanced caption presentation on L2 learners' collocation learning, using eye-tracking technology to monitor visual attention and analyze cognitive processing during video viewing.

A total of sixty participants were divided into two experimental groups, each watching the same video twice under different caption conditions: one group watched English captions with enhanced target collocations first and non-captions second, while the other group watched non-captions first followed by the captioned video with enhanced target collocations. A pretest (before the experiment), comprehension questions (during the experiment), and an immediate posttest (after the experiment) were administered to access their prior knowledge and subsequent learning. Participants' eye movements were tracked with Tobii Pro Lab (version 1.61) and three eye-tracking measures (fixation counts, number of visits, and total fixation duration) were collected and analyzed. The results indicated significant differences in learners' visual attention and collocation learning based on the order of caption presentation. That is, participants in the group who watched the non-captions first followed by the English captions with enhanced collocations outperformed than another group who watched the English captions containing enhanced collocations first and non-captions second in terms of comprehension questions scores and immediate collocation posttest scores. Furthermore, significance differences between the two groups on three eye-tracking measures revealed that the group who watched the non-captions first followed by the English captions featuring enhanced

collocations leads to significantly higher attention allocation and encourages more active cognitive processing of the target collocations. Additionally, the number of visits to the target collocations was the only one that showed a significant positive correlation with the posttest performance, while fixation counts and total fixation duration found no significant importance related to posttest scores.

This study enriches the research on the order of captions in language learning by demonstrating how the order of captions with enhanced collocations affects learners' attention, video comprehension, and collocation learning. The findings have significant implications for the design of instructional videos and multimedia learning materials, indicating that careful consideration of the order of caption presentation can enhance language learning outcomes.

국문초록

성신여자대학교
영어영문학과 우위멍

제2언어(L2) 학습에서 멀티미디어 자료의 활용이 증가함에 따라, 자막이 연어(collocation) 학습을 촉진하는 데 미치는 역할에 대한 관심이 높아지고 있다. 그러나 자막이 언어 학습에 미치는 일반적인 이점에 대한 연구는 많이 이루어졌지만, 자막 제시 순서가 학습에 미치는 영향에 대해서는 거의 알려져 있지 않다. 본 연구는 자막 제시 순서가 L2 학습자의 연어 학습에 미치는 영향을 탐색하고자 하였으며, 실험 중 시청자의 시각적 주의와 인지적 처리를 분석하기 위해 시선추적(eye-tracking) 기술을 활용하였다.

총 60명의 참가자는 두 실험 집단으로 나뉘었으며, 동일한 영상을 서로 다른 자막 조건으로 두 번 시청하였다. 한 집단은 먼저 영어 자막(강조된 목표 연어 포함)을 시청한 후 자막 없이 시청하였고, 다른 집단은 자막 없이 먼저 시청한 후 영어 자막(강조된 연어 포함)으로 시청하였다. 실험 전 사전 테스트(pretest), 실험 중 이해도 질문, 실험 직후 사후 테스트(posttest)를 실시하여 참가자들의 배경지식과 학습 성과를 평가하였다. 참가자의 시선 데이터는 Tobii Pro Lab (버전 1.61)을 통해 수집하였고, 시선 고정 수(fixation counts), 방문 횟수(number of visits), 총 시선 고정 시간(total fixation duration)의 세 가지 지표를 분석하였다.

연구 결과, 자막 제시 순서에 따라 학습자의 시각적 주의 및 연어 학습에 유의미한 차이가 나타났다. 특히 자막 없이 먼저 시청하고 난 후 강조된 연어가 포함된 영어 자막을 본 집단이, 그 반대 순서를 따른 집단보다 이해도 질문과 연어 사후 테스트 모두에서 더 우수한 성과를 보였다. 또한 세 가지 시선 추적 지표 모두에서 두 집단 간 유의미한 차이가 확인되었으며, 자막 없이 먼저 시청한 후 강조된 연어 자막을 시청한 집단이 목표 연어에 더 많은 주의를 기울였고, 보다 적극적인 인지 처리를 수행한 것으로 나타났다. 이와 더불어, 목표 연어에 대한 '방문 횟수'만이 사후 테스트 성과와 유의미한 정적 상관관계를 나타냈으며, 고정 수 및 총 고정 시간은 유의미한 관련성을 보이지 않았다.

본 연구는 자막 제시 순서가 L2 학습자의 시선 집중, 영상 이해 및 연어 학습에 어떠한 영향을 미치는지를 실증적으로 제시함으로써, 외국어 학습 맥락에서 자막 순서에 대한 기존 연구를 심화시켰다. 본 연구 결과는 교육용 영상 및 멀티미디어 학습 자료의 효과적인 설계에 실질적인 시사점을 제공하며, 자막 제시 순서에 대한 세심한 고려가 언어 학습 성과를 높이는 데 기여할 수 있음을 시사한다.