Abstract: Subtitles play a crucial role in facilitating the understanding of visual content when watching films and television programs. In this study, we propose a method for presenting subtitles in a way that considers cognitive load when viewing video content in a non-native language. Subtitles are generally displayed at the bottom of the screen, which causes frequent eye focus switching between subtitles and video, increasing the cognitive load. In our proposed method, we focused on the position, display time, and amount of information contained in the subtitles to reduce the cognitive load and to avoid disturbing the viewer’s concentration. We conducted two experiments to investigate the effects of our proposed subtitle method on gaze distribution, comprehension, and cognitive load during English-language video viewing. Twelve non-native English-speaking subjects participated in the first experiment. The results show that participants’ gazes were more focused around the center of the screen when using our proposed subtitles compared to regular subtitles. Comprehension levels recorded using LightSub were similar, but slightly inferior to those recorded using regular subtitles. However, it was confirmed that most of the participants were viewing the video with a higher cognitive load using the proposed subtitle method. In the second experiment, we investigated subtitles considering connected speech form in English with 18 non-native English speakers. The results revealed that the proposed method, considering connected speech form, demonstrated an improvement in cognitive load during video viewing but it remained higher than that of regular subtitles.

Keywords: subtitles; gaze; cognitive load; user interface

1. Introduction

Native-language subtitles play an important role in the understanding of foreign-language video content such as films and television shows. It has been reported that native-language subtitles are beneficial for content understanding when watching foreign-language video content [1]. Some studies have used gaze data analysis for subtitles evaluation, and found certain drawbacks to the use of subtitles, such as eye strain caused by switching focus between subtitles and visuals, and the potential to miss important scenes in the video by focusing too much on subtitles [2]. Therefore, it is desirable to reduce cognitive load for users when watching subtitled videos, allowing them to further concentrate on the visuals. The guidelines for subtitles [3] have traditionally advised that viewers prefer subtitles to be located at the bottom of the screen. However, these guidelines acknowledge that different placements might be necessary to prevent critical information from being obscured, and that it is crucial to avoid obstructing the speaker’s mouth.

In response to this issue, speaker-following subtitles [4] have been proposed, which display subtitles near the speaker in the video. Compared to traditional subtitles displayed at the bottom of the screen, speaker-following subtitles were found to concentrate the viewer’s gaze near the center of the screen, resulting in greater attention to the speaker rather than the subtitles. However, it was reported that locating the text on the screen was
more difficult with dynamic subtitles that changed position. This problem can potentially be solved by fixing the subtitles in a position close to the speaker without dynamically changing their position. There is also the Rapid Serial Visual Presentation (RSVP) [5] technique, which is a method for bypassing eye movements during reading. In this method, text (each word or small group of words) is displayed in the same location. This technique has become popular in presenting text on wearable devices with limited screen space [6,7], such as smart glasses and smart watches. However, only a small number of studies have applied this technique to subtitles in video. Furthermore, it has been demonstrated that people tend to read text even when they understand the audio content [8–11], so it is desirable to reduce redundant information for viewers.

As a result, in this study, we propose LightSub, a subtitle presentation method that does not interfere with video concentration, for the particular use case of English-language videos with Japanese subtitles. In our proposed method, subtitles with reduced information and shorter presentation time are displayed centrally on the screen, as opposed to the traditional placement at the bottom of the screen. This approach can be interpreted as an application of applying the concept of RSVP technology to video viewing. We focused on three aspects of subtitle presentation during foreign-language video viewing: subtitle position, subtitle duration, and amount of information conveyed. In our proposed method, subtitles with reduced information and shorter display time than those of regular subtitles were displayed in the center of the screen. Traditionally, regular subtitles are displayed for approximately 1–3 s; in our approach, the display time of each word in the subtitles was set to 300 ms. The amount of information in the subtitles was reduced by displaying them on a word-by-word basis instead of in complete sentences, and selecting the words to be displayed based on their lexical level. Our proposed LightSub subtitles differ from RSVP in terms of both word order and displayed information. These changes are necessary to accurately represent Japanese subtitles that correspond to English words as they appear on the screen.

We conducted two experiments to investigate the effects of our proposed subtitle method on gaze distribution, comprehension, and cognitive load during English-language video viewing. In the first experiment, 12 participants took part, and they were asked to watch videos with three different types of subtitles, corresponding to different difficulty levels. The first type was the regular subtitles that are commonly displayed at the bottom of the screen on a per-sentence basis. The second type was the proposed method, which displays subtitles on a per-word basis at the center of the screen. The third type was a video viewing without subtitles. The experimental results indicate that participants’ gazes were focused at the bottom of the screen with regular subtitles. In contrast, with LightSub, their gazes were more evenly distributed toward the center of the screen. Although video comprehension was lower when using the proposed subtitles compared to regular subtitles, it was significantly higher compared to watching without subtitles. The cognitive load was higher than with regular subtitles, indicating that the participants needed to concentrate more while watching the video.

In the first experiment, it was revealed that the subtitles presented with LightSub were fewer than expected by participants. Furthermore, the results of the comprehension test suggested the possibility that words with English-specific phonetic changes (connected speech form) occurring during video viewing were not well perceived.

Therefore, in the second experiment, considering English phonetic changes, we examined and evaluated subtitles with increased information content. A total of 18 participants watched videos, and we investigated the impact on gaze distribution, comprehension, and cognitive load during video viewing. The results showed that comprehension was comparable to regular subtitles. Although cognitive load was higher than with regular subtitles, there was no significant difference, indicating a successful improvement in cognitive load.

Our main contributions can be summarized as follows:

- We propose LightSub, a subtitle presentation method where reduced information and shorter presentation time are centrally displayed. In addition to the display time and
position (spatial) considered in previous studies, we also compared the amount of information with LightSub and regular subtitles.

- The experimental results show that, with the proposed LightSub subtitles, viewers can maintain the same level of video comprehension as regular subtitles even when the amount of information is reduced to approximately 30% of that of the regular subtitles. Further, it is suggested that viewers could concentrate on the scene more than with regular subtitles.

The structure of this paper is as follows. In Section 2, we present related work. Section 3 describes our proposed method. Section 4 describes the first experiment and Section 5 shows the results. Section 6 shows the second experiment and Section 7 presents the results. Section 8 discusses the effect of the proposed method on video viewing. Section 9 presents the drawn conclusions.

2. Related Work

In this section, we describe existing work related to our research. First, we explain English-language learning and extensive viewing; this is followed by a discussion of the role of subtitles in video viewing and subtitle presentation methods.

2.1. English-Language Learning and Extensive Viewing

Amidst the development of the Internet and social networking services (SNSs), the ability to comprehend multiple languages has become crucial for effective communication among individuals whose native languages differ. In the course of this growing international interaction, the acquisition of a second language (L2) has gained prominence. English, being a global lingua franca, is particularly emphasized and frequently utilized on online learning platforms [12]. Furthermore, the population of English as a second language (ESL) learners has been increasing among L2 learners, highlighting the growing necessity of learning English as a second language in recent years. One of the methods for learning English as a second language is known as extensive viewing [13–15]. Extensive viewing involves language learning primarily through visual materials such as dramas and animations. It is defined as "prolonged exposure to easily understandable and entertaining target language materials over an extended period" [16]. Compared to other forms of media like books or radio, visual and auditory stimuli from audiovisual materials, such as films, have been shown to enhance vocabulary acquisition more effectively. Extensive viewing offers various advantages, contributing to the enhancement of learners' speaking skills [17,18], the expansion of vocabulary [19–21], and the exposure to real-life language usage, such as facial expressions and intonation [22,23]. Furthermore, ESL (English as a second language) learners in extensive viewing are shown to prioritize a general understanding and enjoyment of the content over comprehending every detail of the audiovisual material [24]. Additionally, watching personally selected engaging videos has been linked to sustaining learning motivation [25,26], and subtitles have been found to assist in content comprehension [27,28]. Due to these benefits, extensive viewing has gained attention as a crucial learning method for ESL learners [29]. However, many ESL learners find it challenging to comprehend video content at the same level as native speakers, requiring a certain level of language proficiency for content comprehension. If learners cannot understand the content of video materials, continuing to watch becomes difficult. Effectively utilizing video materials is a significant challenge for learners, and the use of subtitles is considered one possible method. In this paper, we assume the viewing of English-language videos with Japanese subtitles and conduct experiments to explore this approach.

2.2. Role of Subtitles in Video Viewing

In various audiovisual works such as movies and dramas, subtitles play a crucial role in aiding the comprehension of content [27,28,30]. Originally used for individuals with hearing impairments to enjoy television and movies, subtitles have been employed in education, particularly in ESL (English as a second language) learning, since the 1980s.
The effectiveness of instructional materials utilizing subtitles has been confirmed through numerous studies [31–33]. Subtitles not only serve as a means to verify the match between visual and auditory information [31] but also play a role in aiding information processing, supplementing audio information, and disappearing in the process [32], making them effective in the language understanding process. Moreover, subtitles are effective for auditory comprehension regardless of the individual’s English proficiency [34,35]. However, relying solely on subtitle assistance and comprehending the content of the video after understanding the language to some extent are considered different. When learners use subtitles in their native language, it is believed that if the translation accuracy is high, the comprehension of content approaches perfection. However, for learners, using subtitles for all content in video materials does not necessarily lead to understanding and mastering the language. From the perspective of video viewing, it is desirable to prioritize language information obtained from hearing for content comprehension, considering subtitles, which provide visual information, as a supplementary role. Therefore, in this paper, we investigate subtitle presentation methods during video viewing to ensure they do not hinder concentration on the video. There is also a significant body of research on multimedia. Mayer et al. proposed a cognitive theory of multimedia learning [36]. It has been demonstrated that combining visual and auditory information enhances learning outcomes compared to receiving only one type of information [37]. However, the phenomenon known as the redundancy effect [38] suggests that an excessive increase in information can actually decrease learning effectiveness. Moreover, during video viewing, there is a tendency for individuals to read subtitles even when they understand the audio content [8]. Therefore, it is considered desirable to reduce unnecessary information for viewers. In this paper, we investigate subtitles with reduced information compared to regular subtitles.

2.3. Research on Subtitles Display Time and Information Content

Extensive research into subtitle presentation methods has been conducted in recent years. It has been demonstrated that extended text imposes a cognitive load on users, leading to a decrease in comprehension [39]. The Rapid Serial Visual Presentation (RSVP) method, proposed by Forster [40], involves rapidly presenting text at a fixed location. This method reduces eye movements from word to word since the text is consistently presented in the same location [41,42]. RSVP is particularly effective when the amount of information presented at once is limited, as it presents either single words or short phrases. Regarding presentation speed, studies have shown that effective comprehension can be achieved at 200 words per minute (WPM) when reading text using RSVP [43]. However, excessively high presentation speeds can lead to decreased comprehension and increased cognitive load [44]. Rzayev et al. [6] investigated text positioning and presentation methods using smart glasses for reading. The study reported higher text comprehension under RSVP conditions compared to methods where text was presented on a per-sentence basis when reading in a seated position. In a study by Erin et al. [7], the impact of RSVP on reading using a smartwatch was examined. Reading under RSVP conditions maintained comparable comprehension levels to conventional methods but allowed for quicker reading times. However, survey results indicated a preference for conventional reading methods over RSVP. Therefore, considering cognitive load, a subtitle presentation method that minimizes stress and aligns with the viewer’s experience is necessary. Applying RSVP to subtitles during video viewing, taking advantage of its characteristics of presenting limited information at once in a short time, is deemed effective in promoting focused video viewing.

2.4. Research on Subtitles Presentation Position

In recent years, there has been an increase in research that explores the non-traditional positioning of subtitles [45–47]. Research on subtitle presentation position encompasses various approaches, including methods that dynamically alter the subtitle’s location [48]. Kurzhals et al. [4] investigated speaker-following subtitles, a method placing subtitles near
speakers in the video. This approach can be seen as applying Fitts’ law [49,50] to subtitles and video. In speaker-following subtitles, a viewer’s gaze was reported to concentrate near the center of the screen, closer to the speaker than traditional subtitles displayed at the bottom of the screen. However, the dynamic movement of subtitles made it more challenging for viewers to locate the text compared to regular subtitles. In the study by Rzayev et al. [6], which employed RSVP, it was found that placing text in the center resulted in significantly higher comprehension and lower cognitive load than positioning it in the upper right. Locating subtitles is a challenge in speaker-following subtitles, and this issue can potentially be addressed by fixing the subtitles relatively close to the center of the screen. It is also considered challenging to apply Rapid Serial Visual Presentation (RSVP), which is presented momentarily, to speaker-following subtitles due to the dynamic movement of the subtitle’s position. In this paper, we investigate subtitles with a fixed position in the center of the screen to enhance video concentration.

3. Proposed Method

As mentioned in Section 2, there are studies focusing on the display time of subtitles and studies focusing on the presentation position of subtitles. Our research integrates these findings and proposes a method that varies two parameters: the display time and the position of the subtitles. Additionally, this research focuses on the amount of information in the subtitles and examines its impact on eye movement, cognitive load, and comprehension. This aspect of investigating the influence of information quantity constitutes the academic contribution of this study.

In this section, we present LightSub, a subtitle presentation method with reduced information and shorter presentation time, which displays subtitles centrally on the screen. To select the vocabulary to display, we used New Word Level Checker (https://nwlc.pythonanywhere.com/ (accessed on 20 September 2022)). New Word Level Checker is a web application provided for Japanese English language learners, which classifies the level of vocabulary contained in the input English text.

This application is used to evaluate the difficulty level of English teaching materials by teachers or to investigate the vocabulary coverage rate of English text written by students [51]. The vocabulary lists included in New Word Level Checker are New JACET8000, JET2000, SVL12000, and others. For this study, we used the vocabulary list from CEFR-J. CEFR-J is a framework developed to apply that CEFR (Common European Framework of Reference for Languages) [52] to English education in Japan. CEFR is an international standard for measuring the proficiency of foreign language use, and is classified into six levels (A1, A2, B1, B2, C1, and C2). A1 level is considered the most basic level, while C2 is considered equivalent to a native level. The words contained in the experimental videos used to evaluate our system were classified using New Word Level Checker, and approximately 60–70 % of the words were classified in the A1 level, with no words falling in the C1 or C2 levels.

In this study, we define the difficulty of vocabulary as indicative of the importance of information. We excluded the most basic A1 level words (easily understandable vocabulary) and selected A2-level and above words (A2, B1, and B2) for display. In A2 level, language learners are expected to be able to understand and use common everyday expressions, basic phrases, and simple sentences related to personal and familiar topics, such as shopping, family, hobbies, and work, but are not expected to understand foreign language media such as films or radio broadcasts. This selection process was carried out manually. As a result, the amount of information of subtitles has been reduced. By displaying only the vocabulary assumed to be necessary for understanding the content of the video, we considered that it would be possible to reduce the amount of information while maintaining the same level of comprehension as regular subtitles.

We created subtitles using Vrew (https://vrew.voyagerx.com/en/ (accessed on 5 October 2022)), a video editing software. There is no standardized method for appropriate subtitle fonts and colors, as they are highly dependent on the country, broadcaster, and
device. In this study, we used Koruri Bold font with white color for the displayed subtitles. Subtitles were placed manually for each of the regular subtitles and the proposed LightSub subtitles. Figure 1 shows a comparison of regular subtitles and proposed subtitles. Due to copyright reasons, the images of the works themselves are not shown, and similar computer-generated images are shown instead. On the left side of Figure 1, subtitles are displayed as sentences in the bottom of the screen, and on the right side of Figure 1, subtitles are displayed as words in the center of the screen. Presenting keywords in the center of the screen increases the likelihood that these keywords will be recognized in central vision without the need for gaze shifts. Additionally, by reducing the amount of textual information and transforming the experience from reading text to viewing text, we aimed to minimize the disruption to video watching. Although displaying text in the center raises the concern of occluding the video itself, the background video is constantly moving, which means the occluded area is always changing. Therefore, the likelihood of losing background video information is low. The proposed LightSub subtitles are displayed in the center of the screen for a short duration with reduced information compared to regular subtitles. These subtitles disappear from the screen after 300 ms after they are displayed. Subtitles are not displayed at regular intervals because they only show vocabulary that is considered difficult to understand. The proposed method not only avoids interfering with visual concentration but is also considered to play a supportive role in aiding viewers' understanding of the video.

Table 1 shows a comparison of the amount of information contained in the subtitles created for this study. Regarding the information in Table 1, the data acquisition process involved several steps. Firstly, we prepared English scripts for the videos used in the experiment. Next, we selected target vocabulary to be displayed using a tool called New Word Level Checker. Subsequently, the selected vocabulary was translated into Japanese. Finally, we calculated the total number of characters in the Japanese subtitles used in the videos. ID represents the number assigned to the videos used in the experiment. The “Information ratio” column displays the ratio of information contained in the LightSub to the information content contained in the corresponding regular subtitles. The number of characters in the proposed subtitles is approximately 30% less than in the regular subtitles.

Table 2 shows the number of subtitle characters displayed over time in the video, calculated by dividing the total number of displayed characters by the number of times subtitles were displayed on the screen. The number of times the subtitles were displayed...
equals the number of subtitles selected by the New Word Level Checker. For instance, in the case of video 1 with LightSub, the total Japanese subtitle characters displayed in LightSub were 107 characters, and the presentation count (number of target subtitles selected by NWLC) was 33, resulting in approximately 3.2 characters per Japanese subtitle presentation. With the proposed method, approximately three characters, which is less than half the number of characters in regular subtitles, are displayed at once.

Table 1. Comparison of the amount of information between proposed subtitles and regular subtitles.

<table>
<thead>
<tr>
<th>ID</th>
<th>Regular Subtitles</th>
<th>LightSub</th>
<th>Information Ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>292</td>
<td>107</td>
<td>36.6</td>
</tr>
<tr>
<td>2</td>
<td>306</td>
<td>77</td>
<td>25.2</td>
</tr>
<tr>
<td>3</td>
<td>288</td>
<td>81</td>
<td>28.1</td>
</tr>
</tbody>
</table>

Table 2. Number of characters displayed at once.

<table>
<thead>
<tr>
<th>ID</th>
<th>Regular Subtitles</th>
<th>LightSub</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.7</td>
<td>3.2</td>
</tr>
<tr>
<td>2</td>
<td>8.7</td>
<td>3.1</td>
</tr>
<tr>
<td>3</td>
<td>7.6</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Hypotheses

We hypothesized that the proposed subtitle method, which focuses on position, display time, and amount of information, would affect gaze distribution, comprehension, and cognitive load during English-language video viewing.

Hypothesis 1. The concentration level during video viewing is higher for the proposed LightSub subtitles compared to the regular subtitles.

It is expected that displaying subtitles in the center of the screen may improve viewer’s concentration on the video as compared to the displayed at the bottom of the screen because centrally placed subtitles are closer to the speaker in the video.

Hypothesis 2. The comprehension level for a given video is comparable for regular subtitles and proposed LightSub subtitles.

Although the amount of information is reduced to compared to regular subtitles by selecting words according to vocabulary level, viewers can maintain the same level of understanding as with regular subtitles.

Hypothesis 3. The cognitive load during video viewing is lower with the proposed LightSub subtitles than with regular subtitles.

It is assumed that viewers watch the video with lower cognitive load compared to regular subtitles by displaying subtitles with a shorter duration in the center of the screen.

4. Experiment 1

We conducted an experiment to investigate the effects of LightSub on gaze distribution, comprehension, and cognitive load during English-language video viewing. A total of 12 participants (10 males and 2 females, with ages ranging from 22 to 25 years, with an average age of 23.2 years) took part in this experiment. All participants were native Japanese speakers and learning English as a second language. The nature of the research, as well as the objectives and procedures of the experiments, were explained to all participants. The research results were utilized solely for the purpose of the study, and participant
information was anonymized and carefully managed. Participants were informed about the consideration given to information management, and they were made aware that they could interrupt the experiment at any time. After consenting to these conditions, participants were asked to sign the consent form for their participation in the experiment. Participants watched English-language videos and answered questions regarding the content of the video. Three subtitle presentation conditions were set. The first condition was regular subtitles displayed at the bottom of the screen on per-sentence basis. The second condition was the proposed subtitles displayed at the center of the screen on per-word basis. The third condition was without subtitles.

We assessed subjective cognitive load using the NASA-TLX [53], collected gaze data with a Tobii Pro Nano (https://www.tobii.com/products/eye-trackers/screen-based/tobii-pro-nano (accessed on 18 January 2023)), screen-based eye tracker, and created heat map to show the gaze distribution during video viewing. Participants evaluated each item of the NASA-TLX for a task involving watching English-language videos under each condition and answering comprehension tests. We used video clips from Friends (https://www.warnerbros.com/tv/friends (accessed on 5 October 2022)), an American comedy show for the experiment (Table 3). We selected clips from scenes in which two or more people were conversing. We chose videos with roughly the same WPM (words per minute) and length of 2–2.5 min.

Table 3. Video clips used in the experiment 1.

<table>
<thead>
<tr>
<th>ID</th>
<th>Video</th>
<th>Duration (m:s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rachel Falls off the Balcony</td>
<td>2:25</td>
</tr>
<tr>
<td>2</td>
<td>Phoebe Gives Monica a Haircut</td>
<td>2:03</td>
</tr>
<tr>
<td>3</td>
<td>Monica Won’t Be Chandler’s Girlfriend</td>
<td>2:17</td>
</tr>
</tbody>
</table>

Procedure

Figure 2 shows the experimental scene. First, participants were seated in a chair, and received an explanation about the video to be viewed and the subtitles to be displayed. The same explanation was given to each participant to keep the experimental environment consistent. Next, participants calibrated the eye tracker and confirmed that it was capable of measuring gaze data.

Figure 2. Experimental scene: participant is watching the video, and gaze data were measured by Tobii Pro Nano, a screen-based eye tracker.
Participants then watched the videos wearing earphones, and after watching, answered a comprehension test (4 choices, 10 questions total) and the NASA-TLX questionnaire (7-point Likert scale). All participants watched each of the subtitle presentation methods once. The combination of the video and subtitles presentation method was randomized. After completing the tasks, participants were asked to fill out questionnaires about their impressions of the subtitle presentation methods. The comprehension test, the NASA-TLX questionnaire, and the questionnaires were all written in Japanese. Therefore, the participants’ ability to answer these questions was not affected by their English reading level. The full study took approximately 30 min to complete per person.

5. Results: Experiment 1

In this section, we discuss the results of the experiment 1. First, we describe the gaze distribution results, followed by comprehension cognitive load results, and finally the questionnaire results. The gaze distribution data include the between-subject factors; the comprehension level and cognitive load were within-subject factors.

5.1. Gaze Distribution

We created a heat map showing the gaze distribution obtained by the eye tracking data measured by Tobii Pro Nano. Regarding the Tobii Pro Nano, the data sampling rate was set to 60 Hz. As for the spatial resolution of the gaze data, the screen resolution used in our study was 1920 × 1080 pixels. Furthermore, when generating a heat map using Python programming, we configured the resolution to 35 × 35 bins. Each bin corresponds to a section of the screen, effectively dividing it into a grid of 35 columns and 35 rows. As a result, each heat map illustrates the concentration of gaze within these grid segments, offering insights into the distribution of participants’ visual attention across the screen.

Figure 3 shows the distribution of gaze of the participants during video viewing (brighter areas mean the gaze is more concentrated). The participants’ gaze is focused on the center and the bottom of the screen with regular subtitles, but the most gazed at area is the bottom of the screen where the subtitles are displayed. On the other hand, with LightSub, the gaze is focused on the center of the screen and is relatively similar to that without subtitles. Therefore, it is suggested that the participants could concentrate on the scene more when using LightSub compared to regular subtitles.

To quantitatively compare the gaze distribution spread when viewing videos with the proposed LightSub and without subtitles, we binarized the heat map images and quantified the size of the rectangular areas encompassing clusters of dots gathered around the center (Figure 4). The two methods were then compared. For the binarization, although Otsu’s method is commonly used for recognizing markers in images with distinct classes of dark and light dots, applying it to our images resulted in a classification of only the central few dots and the rest. Therefore, we manually set the threshold. The threshold setting method involved applying a common threshold to the heat map images of both methods and adjusting it to exclude scattered dots around the periphery. After that, we set a rectangle...
encompassing the cluster of dots in the center and compared the sizes. The proposed method covers an area of 11 dots wide by 9 dots high, whereas the condition without subtitles covers an area of 10 dots wide by 7 dots high. This suggests that the proposed method covers a slightly larger area compared to the no-subtitle condition, although the density of bright dots within the rectangle differs.

Figure 4. The process and results for quantifying and comparing the spread of gaze distribution.

5.2. Comprehension Level

The results of the comprehension test are shown in Figure 5a,b. The triangle in the figure represents the mean, and the black line represents the median of the comprehension score. The circle points indicate outliers. * indicates a significant difference. (p-value was 0.05). The comprehension test scores for each video were higher for video 1 (M = 8.5), video 3 (M = 7.7), and video 2 (M = 7.3), in that order. A one-way analysis of variance (ANOVA) revealed no significant difference in comprehension between the videos (F(2, 33) = 1.65, p > 0.05, η² = 0.102). Therefore, adjusting for difficulty by video was not considered to be problematic. The comprehension test scores for each presentation method were higher for regular subtitles (M = 8.9), proposed subtitles (M = 8.1), and without subtitles (M = 6.4), in that order. A one-way ANOVA revealed a significant difference in comprehension between the presentation methods (F(2, 33) = 11.96, p < 0.01, η² = 0.407). We calculated adjusted significance level of α = 0.0167 using the Bonferroni correction (uncorrected p = 0.05). The result of Bonferroni’s multiple comparison test showed that significant differences were observed between the conditions of regular subtitles and without subtitles, and between the proposed subtitles and without subtitles. The comprehension score of proposed subtitles was significantly higher than that without subtitles. Although the comprehension score of the proposed LightSub subtitles was lower than that of regular subtitles, there was no significant difference between the conditions of regular subtitles and proposed subtitles, suggesting that proposed subtitles can maintain the similar level of comprehension as regular subtitles.
Figure 5. The result of comprehension test. The comprehension test consists of a total of 10 questions, each with four multiple-choice options. The test is scored on a scale of one point per question, with a total of 10 points available for all correct answers and 0 points awarded for all incorrect answers. (a) Comprehension score for each video. (b) Comprehension score for each subtitle method.

5.3. Cognitive Load

The results of the cognitive load evaluation are shown in Figure 6a–c. The triangle in the figure represents the mean, and the black line represents the median of evaluated score. The circle points indicate outlier. * indicates a significant difference. (p-value was 0.05.) Figure 6a shows mental demand, where higher values indicate that more mental and perceptual activity was required, such as watching, thinking, and memorizing. Participants’ subjective mental demand was in the following order: without subtitles (M = 5.9), LightSub (M = 5.0), and regular subtitles (M = 3.3). One-way ANOVA revealed a significant difference in mental demand between presentation methods (F(2, 33) = 9.15, p < 0.01, \( \eta^2 = 0.386 \)). We calculated adjusted significance level of \( \alpha = 0.0167 \) using the Bonferroni correction (uncorrected \( p = 0.05 \)). The result of Bonferroni’s multiple comparison test showed that significant differences were observed between the conditions of regular subtitles and LightSub, and between the regular subtitles and without subtitles. The proposed subtitles resulted in a higher perceived mental demand during the tasks compared to regular subtitles, suggesting that participants needed to concentrate more on watching the video.

Figure 6b shows effort, where higher values indicate that more effort was required to complete the tasks of watching the English-language video and answering the comprehension test. Participants’ subjective effort was in the order of the proposed LightSub (M = 4.7), without subtitles (M = 4.5), and regular subtitles (M = 2.8). One-way ANOVA revealed a significant difference in effort expended on the task between presentation methods (F(2, 33) = 6.56, p < 0.01, \( \eta^2 = 0.213 \)). The result of Bonferroni’s multiple comparison test showed that significant differences were observed between the conditions of regular subtitles and LightSub, and between the regular subtitles and without subtitles. Participants had difficulty in accomplishing the tasks with proposed subtitles compared to regular subtitles.

Figure 6c shows frustration, where higher values indicate that participants felt high level of stress during the tasks. Participants’ subjective frustration was in the order of without subtitles (M = 3.9), proposed LightSub (M = 3.8), and regular subtitles (M = 2.3). One-way ANOVA revealed a significant difference in frustration on the task between presentation methods (F(2, 33) = 5.59, p < 0.05, \( \eta^2 = 0.216 \)). The results of Bonferroni’s multiple comparison test showed significant differences between the conditions: between regular subtitles and the proposed LightSub, and between the regular subtitles and the without subtitles condition. Participants experienced a considerable amount of frustration during the tasks with the proposed LightSub compared to regular subtitles.
Figure 6. Evaluation score of cognitive load for each subtitle presentation method. The evaluation score of cognitive load was obtained using the NASA-TLX, with participants rating each item on a 7-point scale (7: very high; 1: very low). (a) Mental demand: how much mental and perceptual activity was required. (b) Effort: how hard participants had to work to accomplish the tasks. (c) Frustration: the extent to which participants felt stressed and anxious.

5.4. Questionnaire Result

After the experiment, participants were asked “Which presentation method used in the experiment do you prefer to watch English films or other content with?” Nine participants chose regular subtitles, two participants chose the proposed LightSub subtitles, and one participant chose without subtitles. Some participants reported that they had difficulty accepting the proposed LighSub subtitles when watching English-language video content because they were accustomed to regular subtitles. They also found it difficult to comprehend the content of the video and felt that more information would make it easier to understand. Furthermore, some participants who chose the regular subtitles stated that they needed to make more cognitive effort because they could not predict when the subtitles would appear with the proposed subtitles condition. On the other hand, participants who chose the proposed LightSub subtitles commented that it was good that the subtitles only assisted with difficult words because they tend to watch when subtitles are displayed.

6. Experiment 2

In the results of Experiment 1, many participants experienced high cognitive load while watching the video with the proposed LightSub. One contributing factor to cognitive load was the short display time of the subtitles. Moreover, the proposed LightSub subtitle received feedback from participants, indicating that the information conveyed by the subtitles was less than anticipated. Additionally, the analysis of comprehension tests suggested that there was a high likelihood of participants failing to perceive words, especially connected speech in English during video viewing, which could be a significant contributing factor to inaccuracies in their responses. To alleviate cognitive load, it was deemed necessary to consider and augment the information by addressing the connected speech. Consequently, a method involving the modification of subtitle information and presentation duration was contemplated.

This section will commence by discussing preliminary experiment, followed by a description of experiment 2 and its outcomes.

6.1. Preliminary Experiment

The purpose of the preliminary experiment was to verify the appropriate amount of information in the proposed subtitles. In experiment 2, an investigation was carried out on the impact of subtitle presentation methods that increase information during video viewing. However, in the preliminary experiment, the focus was on examining the validity of the information in the subtitles.

As a result of the analysis in experiment 1, it was suggested that the cause of errors in the comprehension test might be the occurrence of connected speech form in English. There-
fore, in the preliminary experiment, attention was directed to words with connected form, and a simplified experiment will be conducted using subtitles with increased information. Regarding English audio variations, Carley et al. [54] defined connected speech form as “A pronunciation of a word influenced by the sounds in surrounding morphemes, syllables, or words” in 2019, highlighting two examples of variations found in General American pronunciation: elision and assimilation. In their 2021 work [55], the authors introduced liaison, where a sound not present in the citation form of a word is inserted in the connected speech form; while there is no significant difference in definitions across these sources, the classification of the phenomenon and the terminology used are not entirely consistent. In this study, we will address the three aforementioned connected speech forms; Table 4 illustrates the rules.

Table 4. English connected speech forms considered in this study

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liaison</td>
<td>The linking of the final consonant of a word with the vowel of the next word</td>
<td>take on, afraid of</td>
</tr>
<tr>
<td>Assimilation</td>
<td>Linking occurs, and the sound changes to a different one</td>
<td>get you, right away</td>
</tr>
<tr>
<td>Elision</td>
<td>Overlapping consonants, resulting in the omission of the expected sound</td>
<td>some more, get tired</td>
</tr>
</tbody>
</table>

6.1.1. Overview of Preliminary Experiment

The experiment involved 12 participants (11 males, 1 female). As mentioned in Section 6.1, the aim of this preliminary experiment was to validate the appropriate amount of information in subtitles. To explore this, three types of subtitles with varying information levels were prepared as candidates. In ascending order of information content, the subtitles included those applying the CEFR rules used in experiment 1, subtitles applying only the connected speech form rules outlined in Table 4, and subtitles applying both CEFR and connected speech form. The preliminary experiment utilized the same footage from the TV show *Friends* to conduct a simplified analysis, creating three video clips approximately 30 s long.

The experimental procedure is outlined as follows: Participants viewed three 30 s videos and subsequently responded to a questionnaire. Subtitles were selected based on CEFR, connected speech form, or a combination of both, providing a comprehensive exploration of subtitle information levels.

The questionnaire consists of the following questions.

- To what extent do you subjectively believe you understood the content of the conversation after watching the video?
- Did you feel that the displayed subtitles were insufficient (too few) while watching the video?
- Did you perceive the displayed subtitles as redundant (unnecessarily excessive) while watching the video?
- When watching foreign-language films in the future, which subtitle presentation method do you prefer?

These questions aim to gather insights on participants’ subjective comprehension, perception of subtitle quantity, feelings about subtitle redundancy, and their preferences for subtitle presentation methods in future instances of watching foreign-language films or similar content.

6.1.2. Results of Preliminary Experiment

The results of the preliminary experiment are presented in Figure 7 through Figure 8b. Figure 7 illustrates the subjective comprehension levels of conversation content for each subtitle presentation method. In subtitles selected using the conventional CEFR rules, the majority of participants had comprehension levels ranging from 20% to 40%, with the highest comprehension reaching only 40% to 60%. On the other hand, for subtitles based on connected speech rules and CEFR and connected speech rules, the graph is primarily
distributed in the 40% to 80% range, suggesting a higher comprehension level compared to CEFR rules alone.

![Figure 7. Subjective video comprehension level.](image)

![Figure 8. How participants perceived the amount of information in subtitles. (a) CEFR condition: Did you perceive the displayed subtitles as insufficient? (b) Connected speech form and CEFR condition: Did you perceive the displayed subtitles as redundant?](image)

The results regarding subtitle information levels are depicted in Figure 8a,b. Figure 8a illustrates participants’ perceptions of subtitle scarcity under the CEFR condition. It reveals that, irrespective of the degree, all participants felt that subtitles were insufficient when selected solely based on CEFR rules. Figure 8b represents participants’ perceptions of subtitle redundancy under the CEFR and connected speech condition. One participant responded with “neither agree nor disagree”; meanwhile, the remaining 11 participants, to varying degrees, felt that the subtitles were not redundant.

The survey conducted in the preliminary experiment asked participants about their preferences for subtitle usage when watching foreign-language films in the future, using the subtitles from the experiment (CEFR only, connected speech form only, CEFR and connected speech form). Two participants preferred subtitles under the connected speech form only, while ten participants preferred subtitles under the CEFR and connected speech form condition. No participants chose the CEFR only condition. Impressions from participants included positive feedback, such as “I want to enjoy foreign-language films as content, so the more information, the easier it is to understand the content” and “Understanding the meaning through subtitles made the audio more intelligible.” Overall, many positive opinions were gathered regarding subtitles with increased information.

Based on these results, subtitles selected under the CEFR and connected speech form condition demonstrated both high comprehension levels and non-intrusiveness according
to user perceptions. Consequently, considering the potential benefits, it was decided to
handle this subtitle presentation method in Experiment 2 for further validation.

6.2. Procedure

The experiment enlisted 18 participants aged 21 to 25 (16 males, 2 females). All
participants were provided with an explanation of the research’s nature, along with the
objectives and procedures of the experiments. The utilization of research results was
exclusively for the study’s purpose, and thorough measures were taken to anonymize and
manage participant information. Participants were briefed on the meticulous information
management practices, and they were informed of their ability to interrupt the experiment
at any point. Upon agreeing to these terms, participants were requested to sign the
consent form indicating their willingness to participate in the experiment. Similar to
experiment 1, participants watched English-language videos and subsequently answered
questions related to the video content. The experiment included three conditions: standard
subtitles presented at the bottom of the screen (regular subtitles), subtitles presented in
the center of the screen with both CEFR and connected speech form considerations (LightSub
(connected speech)), and subtitles presented in the center of the screen based on CEFR rules
alone (LightSub).

For the subjective assessment of cognitive load, the NASA-TLX (NASA Task Load
Index) was employed. Participants evaluated each aspect of NASA-TLX for tasks involving
watching English-language videos under each condition and responding to comprehension
tests. Additionally, the System Usability Scale (SUS) was used to assess usability aspects,
evaluating the ease of use for each subtitle presentation method.

The experiment utilized video clips from the American comedy–drama series Friends,
featuring scenes with two or more individuals engaged in conversation (Table 5). The
selected video clips were approximately 3 to 3.5 min long, with relatively similar words-
per-minute (WPM) rates.

Table 5. Video clips used in experiment 2.

<table>
<thead>
<tr>
<th>ID</th>
<th>Video</th>
<th>Duration (m:s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>The Friends Pretend To Like Rachel’s English Trifle</td>
<td>3:38</td>
</tr>
<tr>
<td>B</td>
<td>Ross Had His Teeth Whitened</td>
<td>3:00</td>
</tr>
<tr>
<td>C</td>
<td>Rachel Hires Tag As Her Assistant</td>
<td>3:35</td>
</tr>
</tbody>
</table>

Moreover, Table 6 presents a comparison of the information content of subtitles created
in experiment 2, while Table 7 illustrates a comparison of subtitles presented simultaneously.
Table 6 compares the number of Japanese subtitle characters presented in experiment 2’s
videos for each method, indicating the information content relative to regular subtitles set
at 100. The information content of subtitles applying only CEFR rules (CEFR condition)
is approximately 20% to 30% of regular subtitles, while subtitles applying both CEFR
and connected speech rules (LightSub (connected speech form)) present about 60% of the
information content of regular subtitles. The information content has increased compared
to the CEFR condition, providing a sufficient amount of subtitles that viewers expect while
watching the video.

Table 7 compares the number of subtitles presented simultaneously in each video, cal-
culated by dividing the total number of subtitle characters by the number of presentations.
Under the connected speech condition, approximately 3–4 characters of the subtitles are
presented at once, which is less than half of regular subtitles.
Table 6. Comparison of information ratio of subtitles used in experiment 2 (%).

<table>
<thead>
<tr>
<th>ID</th>
<th>Regular Subtitles</th>
<th>LightSub (Connected Speech)</th>
<th>LightSub</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>100</td>
<td>55.7</td>
<td>34.2</td>
</tr>
<tr>
<td>B</td>
<td>100</td>
<td>59.3</td>
<td>18.4</td>
</tr>
<tr>
<td>C</td>
<td>100</td>
<td>61.3</td>
<td>27.3</td>
</tr>
</tbody>
</table>

Table 7. Number of characters displayed at once.

<table>
<thead>
<tr>
<th>ID</th>
<th>Regular Subtitles</th>
<th>LightSub (Connected Speech)</th>
<th>LightSub</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7.6</td>
<td>3.5</td>
<td>3.4</td>
</tr>
<tr>
<td>B</td>
<td>7.5</td>
<td>3.4</td>
<td>2.4</td>
</tr>
<tr>
<td>C</td>
<td>7.7</td>
<td>3.4</td>
<td>2.5</td>
</tr>
</tbody>
</table>

The experimental procedure is outlined here. Initially, participants were seated and briefed on the video to be watched and the accompanying subtitles. Subsequently, calibration for the eye tracker was conducted to ensure accurate measurement of eye gaze data. Following this, participants, while wearing headphones, watched the video. They then responded to a comprehension test related to the video (11 multiple-choice questions), NASA-TLX (on a 7-point Likert scale), and SUS (on a 5-point Likert scale).

Eye gaze data were collected using the Tobii Pro Nano, and a heat map illustrating gaze distribution was generated. The process of watching the video and providing responses constituted one set, and this was repeated for three different video clips. The combination of the video each participant watched and the displayed subtitles was randomized. Participants concluded the experiment by filling out a questionnaire. The average time required for each participant to complete the experiment was approximately 30 min.

- Regular subtitles: Japanese subtitles are presented at the bottom of the screen. The duration of subtitle presentation varies based on the length of the text, typically ranging from 1 to 3 s.
- LightSub (connected speech): Japanese subtitles, selected based on both CEFR and connected speech rules, are displayed in the center of the screen. The presentation time for each word is set at 500 ms.
- LightSub: Japanese subtitles chosen based on CEFR rules are presented in the center of the screen. The presentation time for each word is set at 300 ms. The CEFR condition replicates the subtitle presentation method proposed in experiment 1.

7. Results: Experiment 2

The results of the evaluation experiment are presented in terms of gaze distribution, video comprehension, cognitive load, and usability. Gaze distribution is analyzed between participants, while video comprehension, cognitive load, and usability are analyzed within participants.

7.1. Gaze Distribution

Figure 9 shows the distribution of gaze of the participants during video viewing. Brighter areas indicate a more concentrated gaze. For regular subtitles, it is observed that gazes concentrate near the center and bottom of the screen. On the other hand, the gaze fixation locations without subtitles are centered. This suggests that important content in the video is likely concentrated in the center. In the proposed method, the gaze distribution being similar to that without subtitles indicates a higher likelihood that attention was directed towards the video rather than the subtitles; while it is difficult to estimate the internal state of concentration, the gaze patterns suggest a similar condition to when participants were focused on the video content.
Additionally, the actual subtitle presentation time (total) in the videos used for experiment 2 is presented in Table 8.

Table 8. Comparison of the total subtitle presentation time.

<table>
<thead>
<tr>
<th>ID</th>
<th>Regular Subtitles</th>
<th>LightSub (Connected Speech)</th>
<th>LightSub</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>144 s (66%)</td>
<td>46 s (21%)</td>
<td>18 s (8%)</td>
</tr>
<tr>
<td>B</td>
<td>123 s (68%)</td>
<td>36 s (20%)</td>
<td>10 s (6%)</td>
</tr>
<tr>
<td>C</td>
<td>143 s (67%)</td>
<td>48 s (22%)</td>
<td>17 s (8%)</td>
</tr>
</tbody>
</table>

The numbers in parentheses in Table 8 represent the proportion of time subtitles are displayed for each method when considering the entire duration of the video as 100%. For instance, in the case of regular subtitles in video 2, subtitles are presented for 66% of the total video duration, while subtitles are presented for 21% of the total video duration in the LightSub (connected speech form) condition. According to this table, the subtitle presentation time in the LightSub (connected speech) condition is approximately 20% of the total video duration, representing a reduction of about one-third compared to regular subtitles. Hence, it is suggested that participants can focus more on the video compared to regular subtitles.

7.2. Comprehension Level

The results of the comprehension test are illustrated in Figure 10a, b. Figure 10a provides a comparison between individual videos, while Figure 10b presents a comparison between methods. In the figures, green triangles represent the mean values, black lines represent the medians, and * indicates a significant difference in comprehension at a 5% significance level. Regarding the comprehension test scores for each video, video C (\(M = 10.11\)) achieved the highest score, followed by video A (\(M = 10.05\)) and video B (\(M = 9.21\)). One-way analysis of variance (ANOVA) indicated no significant differences in comprehension among the videos (\(F(2, 51) = 3.82, p > 0.05, \eta^2 = 0.130\)). Therefore, it is considered that the difficulty level adjustment across videos was appropriate. The comprehension test scores for each method were higher in the following order: regular subtitles (\(M = 10.26\)), LightSub (connected speech) (\(M = 10.0\)), and LightSub (\(M = 9.11\)). One-way analysis of variance (ANOVA) revealed a significant difference in comprehension among the methods (\(F(2, 51) = 6.20, p < 0.01, \eta^2 = 0.344\)). To account for multiple comparisons, the Bonferroni method was applied, yielding an adjusted significance level of \(a = 0.0167\) (pre-adjustment \(p = 0.05\)). Subsequent Bonferroni multiple comparison tests confirmed significant differences between regular subtitles and LightSub, as well as between LightSub (connected speech) and LightSub. The subtitle method considering connected speech demonstrated significantly higher comprehension compared to the CEFR condition. Moreover, there was no significant difference between regular subtitles and the LightSub (connected speech), suggesting that the proposed method maintained a level of comprehension similar to regular subtitles.
Figure 10. The results of the comprehension test. The comprehension test consists of a total of 11 questions, each with 4 multiple-choice options. The test is scored on a scale of 1 point per question, with a total of 11 points available for all correct answers and 0 points awarded for all incorrect answers. (a) Comprehension score for each video. (b) Comprehension score for each subtitle method.

7.3. Cognitive Load

The results of subjective cognitive load during tasks are illustrated in Figure 11a–c. The green triangles represent the mean values, black lines indicate the medians, and * denotes a significant difference at a 5% significance level. Figure 11a portrays the perceived cognitive load, where higher values indicate increased load related to tasks such as viewing, thinking, and memorizing. Participants’ subjective perceived cognitive load ranked in the following order: LightSub ($M = 4.2$), LightSub (connected speech) ($M = 3.7$), and regular subtitles ($M = 2.7$). One-way analysis of variance (ANOVA) revealed a significant difference in perceived cognitive load among the methods ($F(2, 51) = 6.82, p < 0.01, \eta^2 = 0.143$). Applying the Bonferroni method to adjust the significance level ($\alpha = 0.0167$, pre-adjustment $p = 0.05$), subsequent Bonferroni multiple comparison tests showed a significant difference between regular subtitles and the LightSub. Although the cognitive load of LightSub (connected speech) during tasks was slightly higher compared to regular subtitles, no significant difference was observed.

Figure 11b illustrates the extent to which participants exerted effort to accomplish the task, with higher values indicating greater struggles in watching English-language videos and responding to comprehension tests. Participants subjectively reported effort in the following order: LightSub ($M = 4.3$), LightSub (connected speech) ($M = 3.7$), and regular subtitles ($M = 2.94$). One-way analysis of variance (ANOVA) revealed a significant difference in effort exerted among the methods ($F(2, 51) = 5.04, p < 0.05, \eta^2 = 0.119$). Subsequent Bonferroni multiple comparison tests indicated a significant difference between regular subtitles and the LightSub. For LightSub (connected speech), although more effort was required compared to regular subtitles, no significant difference was observed in task accomplishment.

Figure 11c represents the stress participants perceived during tasks, with higher values indicating greater stress. Subjective stress reported by participants ranked in the following order: LightSub ($M = 3.0$), LightSub (connected speech) ($M = 2.8$), and regular subtitles ($M = 2.2$). One-way analysis of variance (ANOVA) did not reveal a significant difference in stress perception among the methods during tasks ($F(2, 51) = 1.71, p > 0.05, \eta^2 = 0.0038$). In this experiment, no significant differences in participants’ subjective stress were observed among the regular subtitle, the LightSub, and the LightSub (connected speech) conditions.
Figure 11. Evaluation scores of cognitive load for each subtitle presentation method. The evaluation score of cognitive load was obtained using the NASA-TLX, with participants rating each item on a 7-point scale (7: very high; 1: very low). (a) Mental demand: how much mental and perceptual activity was required. (b) Effort: how hard participants had to work to accomplish the tasks. (c) Frustration: the extent to which participants felt stressed and anxious.

7.4. Result of System Usability Scale

Figure 12 presents the results of the System Usability Scale (SUS) assessment. The average scores were 83.3 points for regular subtitles, 66.1 points for LightSub (connected speech), and 57.9 points for LightSub. This indicates that regular subtitles were perceived as the most user-friendly. The median scores were 85.0 points for regular subtitles, 72.5 points for the LightSub (connected speech), and 62.5 points for the LightSub. The average score for typical systems is considered to be 68 points, and the LightSub (connected speech) condition slightly fell below this average.

Figure 12. System usability scale evaluation.

Regarding usability evaluation, the average scores for each item in the SUS questionnaire for experiment 2 were calculated. Table 9 presents the average values for positive questions. Each item is rated on a scale of 1 to 5, where higher values indicate better results for positive questions. For the proposed method, all average values are above 3.0, indicating that the usability of subtitles is comparable to regular subtitles. However, a difference was observed in the items “This subtitle presentation method functions smoothly and is well integrated” and “I feel confident in using this subtitle presentation method”. For these items, there was a distinction compared to regular subtitles. It is suggested that the proposed subtitle method may be less familiar to participants, requiring some adaptation for proficient use.
Table 9. Mean scores of SUS for positive questions.

<table>
<thead>
<tr>
<th>ID</th>
<th>Question</th>
<th>Regular</th>
<th>LightSub (Connected Speech)</th>
<th>LightSub</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I would like to use the subtitle presentation method used in this video frequently.</td>
<td>3.4</td>
<td>3.1</td>
<td>2.8</td>
</tr>
<tr>
<td>3</td>
<td>The subtitle presentation method used in this video was easy to use.</td>
<td>3.5</td>
<td>3.3</td>
<td>2.8</td>
</tr>
<tr>
<td>5</td>
<td>I felt that this subtitle presentation method works smoothly and is well-integrated.</td>
<td>4.3</td>
<td>3.7</td>
<td>3.1</td>
</tr>
<tr>
<td>7</td>
<td>Most people would quickly understand how to use this subtitle presentation method.</td>
<td>4.8</td>
<td>4.3</td>
<td>3.9</td>
</tr>
<tr>
<td>9</td>
<td>I am confident that I can master this subtitle presentation method</td>
<td>4.1</td>
<td>3.1</td>
<td>2.9</td>
</tr>
</tbody>
</table>

Table 10 presents the average values for negative questions. Each item is rated on a scale of 1 to 5, where lower values indicate better results for negative questions. Two items that showed significant differences compared to regular subtitles are “There were inconsistencies in this subtitle presentation method” and “A lot of prior knowledge is required before using this subtitle presentation method”. It is suggested that, in the proposed method, participants might have felt discomfort due to subtle nuances arising from the translation of English content into Japanese subtitles. This could be attributed to the inherent challenges of accurately translating English content into Japanese while preserving the intended meaning.

Table 10. Mean scores of SUS for negative questions

<table>
<thead>
<tr>
<th>ID</th>
<th>Question</th>
<th>Regular</th>
<th>LightSub (Connected Speech)</th>
<th>LightSub</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>The subtitle presentation method used in this video was unnecessarily complex.</td>
<td>1.5</td>
<td>1.9</td>
<td>1.9</td>
</tr>
<tr>
<td>4</td>
<td>I think that expert support is needed to use this subtitle presentation method.</td>
<td>1.2</td>
<td>1.9</td>
<td>2.2</td>
</tr>
<tr>
<td>6</td>
<td>I felt that there were inconsistencies in this subtitle presentation method.</td>
<td>1.2</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>8</td>
<td>I found this subtitle presentation method difficult to use.</td>
<td>1.9</td>
<td>2.3</td>
<td>2.9</td>
</tr>
<tr>
<td>10</td>
<td>A lot of prior knowledge is required before using this subtitle presentation method.</td>
<td>1.4</td>
<td>2.3</td>
<td>2.8</td>
</tr>
</tbody>
</table>

7.5. Questionnaire Result

The results of the post-experiment survey are presented here. When asked about their preference for the methods used in the experiment (regular subtitles, LightSub, and LightSub (connected speech)) when watching foreign-language films in the future, eight participants chose regular subtitles, eight participants chose LightSub (connected speech), and two participants chose LightSub.

Participants were further questioned about the reasons behind their choice of subtitle presentation method. For regular subtitles, responses included “Because it’s easy to understand in the format I’m used to” and “With other presentation methods, not everything is explained, so there are parts where I cannot understand the meaning”.

Regarding the LightSub (connected speech), participants mentioned that “The subtitles were simple, covering a reasonable number of words, and there was a balance between understanding spoken English and reading subtitles” and “There were fewer eye movements, and the subtitles complemented exactly where I could not hear”.

For the CEFR condition, responses included the following: “It seems like I can focus on the movie” and “I thought it would be a study in both understanding the content and improving listening skills”.

8. Discussion and Limitation

8.1. Discussion

In experiment 2, the results indicated that LightSub (connected speech), which increased the amount of information, showed significantly higher comprehension compared to the LightSub. Although statistical significance was not observed, there was a reduction in cognitive load. It is inferred that the proposed method successfully achieved its goal by assisting in the comprehension of areas where connected speech occur, leading to improved comprehension and reduced cognitive load.

We conducted a survey of the participants’ TOEIC Listening and Reading test (https://www.iibc-global.org/toeic/test/lr.html (accessed on 10 February 2023)) scores that they have taken
in the past. Table 11 shows the distribution of participants’ TOEIC Listening and Reading test scores in experiment 1 and experiment 2. In experiment 1, the two participants who favored the proposed LightSub subtitles both had TOEIC scores in the 800s, representing the highest scores among the participants. Comparing comprehension test scores, the scores in experiment 2 were relatively higher than those in experiment 1. This could be attributed to the possibility that either the videos or the comprehension test in experiment 2 were relatively easier. However, as shown in Table 11, participants in experiment 2 generally had a reasonable proficiency in English, which could also contribute to the higher scores. Furthermore, the survey results revealed that eight participants expressed their interest in using the proposed method for future activities such as watching foreign-language films. This number, equal to the preference for regular subtitles, marked a significant increase compared to experiment 1, suggesting that this subtitle presentation method is more suitable for users with a certain level of English proficiency.

Table 11. Distribution of participants’ TOEIC Listening and Reading Test Scores.

<table>
<thead>
<tr>
<th>Score Range</th>
<th>Number of Participants (Experiment 1)</th>
<th>Number of Participants (Experiment 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>900–990</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>800–895</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>700–795</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>600–695</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>500–595</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>400–495</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

8.2. Limitations

Our proposed approach presents a number of limitations. Firstly, one limitation is the restricted applicability of the proposed LightSub subtitles. Since the amount of information is less than the regular subtitles, the comprehension level depends on the participants’ English proficiency level. It might be challenging for English beginners to effectively use our proposed subtitles. Additionally, the focus of this method is on Japanese learners studying English as a second language. Subtitles in videos serve a broad purpose, and text information is crucial, especially for individuals with hearing impairments. The reduced information in subtitles might pose difficulties for those with hearing impairments to use this method effectively. Secondly, the impact of the selected video for the experiment is another challenge. In this study, *Friends* was chosen as the experimental video, characterized by multiple conversational scenes and colloquial expressions. Using a different video for the experiment may result in varying effectiveness of the proposed method. Lastly, the manual creation of subtitles is a limitation. The primary objective of this study is to evaluate the impact of each subtitle presentation method on gaze distribution, comprehension, and cognitive load during video viewing. The manual creation of subtitles is a current limitation, and automatic subtitle generation is a potential area for future research.

We also considered the following two reasons for the higher cognitive load of this method compared to the conventional method: One possible reason why the proposed method did not reduce cognitive load compared to conventional subtitle presentation methods may be the selection of video content. In both experiment 1 and experiment 2, the cognitive load was lower with conventional subtitle presentation methods, whereas it was higher without subtitles. This suggests that understanding the content was difficult without subtitles. In fact, many participants mentioned that it was challenging to understand the English audio alone. Therefore, many participants relied on the subtitles for content understanding, and reducing subtitle information may have contributed to an increase in cognitive load. One countermeasure could be to lower the English level of the videos. It would be necessary to compare the cognitive load of the proposed method and conventional methods using content that can be somewhat understood through audio alone.
Another reason why the proposed method did not reduce cognitive load might be because the subtitles were presented in Japanese. The video content had English audio, and understanding it in Japanese while listening to the English audio might have increased cognitive load due to the need to switch between English and Japanese modes. With conventional methods where all subtitles are in Japanese, participants can primarily rely on Japanese for understanding. However, with the proposed method, since only minimal Japanese is presented, maintaining an understanding mode in English might have been necessary. A possible solution could be to present keywords in pairs of English and Japanese. This way, participants could utilize Japanese keywords while being in the mode of listening to English.

Regarding the sample size of the two experiments, the number of participants in this study was 12 for experiment 1 and 18 for experiment 2, which is somewhat small compared to previous studies. To obtain more solid conclusions, future experiments with larger sample sizes are needed.

9. Conclusions

In this study, we proposed LightSub, a subtitle presentation method focusing on the position, duration, and amount of information to reduce the cognitive load and to avoid disturbing the viewer’s concentration. We investigated the effects of our proposed subtitle method on gaze distribution, comprehension, and cognitive load during English-language video viewing. Below, we present the conclusions for the hypotheses set in this study.

Hypothesis 4. The concentration level during video viewing is higher for the proposed LightSub subtitles compared to the regular subtitles.

With the regular subtitles, participants’ gazes mostly focused on the bottom of the screen. On the other hand, with LightSub, gazes are focused on the center of the screen, similarly to when watching without subtitles. It was suggested that participants were able to concentrate more on the scene when the subtitles were fixed in a center position closer to the speaker, compared to the regular subtitles.

Hypothesis 5. The comprehension level for a given video is comparable for regular subtitles and proposed LightSub subtitles.

In experiment 1, the analysis revealed that the comprehension score of proposed LightSub subtitles was significantly higher than that of without subtitles. Although comprehension score of proposed subtitles was lower compared to regular subtitles, there was no significant difference between the conditions of regular subtitles and proposed subtitles. The results suggest that LightSub can maintain the same level of comprehension as regular subtitles despite reducing the amount of information by approximately 30% compared to regular subtitles. In experiment 2, LightSub taking into account connected speech also can maintain the same level of comprehension as regular subtitles even though the amount of information is reduced to approximately 60% of regular subtitles.

Hypothesis 6. The cognitive load during video viewing is lower with the proposed LightSub subtitles than with regular subtitles.

The analysis of cognitive load revealed that with the proposed subtitles, subjective mental demand, effort, and frustration were all higher than with regular subtitles. We attempted to reduce the cognitive load of viewers during video watching by fixing the subtitles in a center position with reduced information and shorter presentation time; however, our proposed method was not effective in reducing the cognitive load.

Furthermore, the subtitles positioned centrally possess the potential to impede the user’s view of the visual content. Modifying the subtitle style, such as making subtitles partially transparent, could minimize disturbance of video content and improve the user
experience. Additionally, it is assumed that the effectiveness of the method may differ depending on the characteristics of the video, not just videos with many colloquial expressions, but also speeches with a formal atmosphere. Considering the various constraints outlined above and the indication from the experiments that the proposed method did not reduce cognitive load compared to regular subtitles, we acknowledge that some aspects of this approach may be deemed inappropriate. This suggests the need to explore alternative solutions or further refine existing approaches to achieve a more effective subtitle presentation method.

In the future, we are considering a method that dynamically changes the level of English words to be excluded and displayed on the screen according to the viewer’s English proficiency.

Author Contributions: Conceptualization, Y.N. (Yuki Nishi), Y.N. (Yugo Nakamura), S.F. and Y.A.; methodology, Y.N. (Yuki Nishi); software, Y.N. (Yuki Nishi); validation, Y.N. (Yuki Nishi), Y.N. (Yugo Nakamura), S.F. and Y.A.; formal analysis, Y.N. (Yuki Nishi); investigation, Y.N. (Yuki Nishi); resources, Y.N. (Yuki Nishi); data curation, Y.N. (Yuki Nishi); writing—original draft preparation, Y.N. (Yuki Nishi); writing—review and editing, Y.N. (Yuki Nishi), Y.N. (Yugo Nakamura), S.F. and Y.A.; visualization, Y.N. (Yuki Nishi); supervision, Y.N. (Yugo Nakamura), S.F. and Y.A.; project administration, Y.N. (Yuki Nishi); funding acquisition, Y.N. (Yugo Nakamura), S.F. and Y.A. All authors have read and agreed to the published version of the manuscript.

Funding: This study was partly supported by MEXT “Innovation Platform for Society 5.0” Program Grant Number JPMXP0518071489.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Kyushu University (ISEE 2023-23 (25 January 2024)).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data collected during the study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References
11. Ross, N.M.; Kowler, E. Eye movements while viewing narrated, captioned, and silent videos. *J. Vis.* 2013, 13, 1. [CrossRef] [PubMed]


21. Masrai, A. Can L2 phonological vocabulary knowledge and listening comprehension be developed through extensive movie viewing? The case of Arab EFL learners. *Int. J. List.* 2020, 34, 54–69. [CrossRef]


23. Renandya, W.A.; Farrell, T.S. ‘Teacher, the tape is too fast!’ Extensive listening in ELT. *ELT J.* 2011, 65, 52–59. [CrossRef]


29. Ivone, F.M.; Renandya, W.A. Extensive listening and viewing in ELT. *Teflin J.* 2019, 30, 237–256. [CrossRef]


49. Fitts, P.M. The information capacity of the human motor system in controlling the amplitude of movement. J. Exp. Psychol. 1954, 47, 381. [CrossRef] [PubMed]


51. Milliner, B. Evaluating the lexical difficulty of teaching materials with NWLC. ELF 2 2022, 2, 49.


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.