Applied Psycholinguistics (2022), **43**, 237–269
doi:10.1017/S0142716421000527

CAMBRIDGE
UNIVERSITY PRESS

ORIGINAL ARTICLE

The impact of audio on the reading of intralingual versus interlingual subtitles: Evidence from eye movements

Sixin Liao[1,3]*, Lili Yu[2,3], Jan-Louis Kruger[1,3] and Erik D. Reichle[2,3]

[1]Department of Linguistics, Macquarie University, Sydney, NSW, Australia, [2]Department of Psychology, Macquarie University, Sydney, NSW, Australia and [3]Macquarie University Centre for Reading, Sydney, NSW, Australia
*Corresponding author. Email: sixin.liao@students.mq.edu.au

(Received 12 January 2021; revised 17 July 2021; accepted 18 October 2021; first published online 03 December 2021)

Abstract
This study investigated how semantically relevant auditory information might affect the reading of subtitles, and if such effects might be modulated by the concurrent video content. Thirty-four native Chinese speakers with English as their second language watched video with English subtitles in six conditions defined by manipulating the nature of the audio (Chinese/L1 audio vs. English/L2 audio vs. no audio) and the presence versus absence of video content. Global eye-movement analyses showed that participants tended to rely less on subtitles with Chinese or English audio than without audio, and the effects of audio were more pronounced in the presence of video presentation. Lexical processing of subtitles was not modulated by the audio. However, Chinese audio, which presumably obviated the need to read the subtitles, resulted in more superficial post-lexical processing of the subtitles relative to either the English or no audio. On the contrary, English audio accentuated post-lexical processing of the subtitles compared with Chinese audio or no audio, indicating that participants might use English audio to support subtitle reading (or vice versa) and thus engaged in deeper processing of the subtitles. These findings suggest that, in multimodal reading situations, eye movements are not only controlled by processing difficulties associated with properties of words (e.g., their frequency and length) but also guided by metacognitive strategies involved in monitoring comprehension and its online modulation by different information sources.

Keywords: audio; eye movements; multimodal integrated-language processing; reading; subtitles

The global dissemination of audio-visual products such as films and educational videos has resulted in widespread use of subtitling as a tool to minimize language barriers and enhance media accessibility (Kruger & Doherty, 2016; Liao et al., 2020). Although subtitle reading has been a topic of growing interest over the past decades, our understanding of the mental processes engaged during the reading of subtitles, and how that differs from the reading of static text, is still limited (Kruger, 2016).

Our limited understanding likely reflects the inherent complexity involved in subtitle reading—that is, reading subtitles needs to be coordinated with other more or less demanding tasks, such as identifying objects from the background video and/or comprehending the linguistic content of the audio, all of which compete for limited attention and are subject to a variety of other perceptual (e.g., visual acuity) and cognitive (e.g., working memory) limitations.

The present study aims at contributing to our understanding of this complicated multimodal reading behavior involved in watching videos with subtitles, by providing more insights into the influence of audio on the reading and comprehension of subtitles. We first start with some relevant theoretical accounts and empirical evidence for text-picture integration, mainly focusing on the *cognitive theory of multimedia learning* by Mayer (2005, 2014) and the *multimodal integrated-language framework* by Liao et al. (2021). We then review what has been learned from previous empirical studies about the influence of audio on the reading of static text and subtitles. After that, we present an eye-tracking experiment that examined the interaction between the auditory and visual systems when watching videos with subtitles. Finally, we discuss how the multimodal integrated-language framework could be extended based on the empirical data from the present study to better understand reading in multimodal contexts, such as reading subtitles in educational videos.

Theoretical frameworks

The sheer complexity of the mental processes that support multimodal reading has motivated attempts to develop theories to explain what transpires in the mind of someone who is, for example, engaged in the watching of a subtitled film. Perhaps because of the complexity of what must be explained, however, these theories provide only high-level descriptions (rather than, e.g., being mathematical or computational models) that expand upon basic principles of cognitive psychology (e.g., Atkinson & Shiffrin, 1968, 1971) to explain the phenomena of interest. For example, the theories described how "streams" of information from two or more sensory modalities or formats (e.g., text vs. images) are encoded and processed to construct abstract representations of whatever content is being understood. As such, although the theories are limited in their utility, they nonetheless provide useful frameworks for both thinking about and making (qualitative) predictions about situations involving multimodal reading.

For example, Mayer's (2005) *cognitive theory of multimodal learning* assumes that: (1) both linguistic and non-linguistic information can be extracted via the visual and/or auditory channels; (2) this information is actively processed in working memory; and (3) the rate of information processing is delimited by attention, working memory capacity, and the encoding/retrieval of information into/from long-term memory. These assumptions are consistent with what is generally known about cognition (Atkinson & Shiffrin, 1968, 1971) and thus allow the theory to, for example, explain why the processing of two information sources within one sensory modality is typically more difficult than processing across two modalities (e.g., visual acuity limitations cause image viewing to be more disruptive to reading than does listening to music).

Similarly, Schnotz's (2005) *integrated text and picture comprehension theory* describes in slightly more detail the subprocesses involved in multimodal reading (e.g., how the pronunciations of printed words are generated via grapheme-to-phoneme rules) and distinguishes between *descriptive* or symbolic media and representations (e.g., written and spoken language) versus *depictive* or analog media and representations (e.g., video images). The theory also distinguishes between an early *surface* or perceptual stage in which visual and auditory information is converted into linguistic and non-linguistic patterns, and a subsequent *deep* or semantic stage in which these patterns undergo descriptive or depictive processing to construct propositional representations of whatever is being conveyed by the media. Despite these additional assumptions, the theory is also largely limited to making qualitative predictions.

Although these two theories and others (e.g., see Cohn, 2016) have advanced our understanding of multimodal reading situations (e.g., see Mayer, 2014), we firmly believe that further progress will benefit from more formal theories (as exemplified in reading research; see Reichle, 2021), and for that reason, we have directed our efforts toward conducting eye-movement experiments that allow "traction" in the development of more computational accounts of multimodal reading. For example, in a previous attempt to examine how concurrent video content and subtitle presentation speed affect comprehension and various online indicators of language processing, Liao et al. (2021) found that simultaneous display of video and subtitle improved viewers' comprehension even with a fast subtitle speed (e.g., 20 characters per second). Global eye-movement analyses revealed that increasing subtitle speed resulted in fewer, shorter fixations and longer saccades on the subtitles. Furthermore, their examination of the word-length effect, the word-frequency effect, and the wrap-up effect provided more detailed information about how the reading of subtitles was modulated by the global task demands due to the presence of video content and/or increasing subtitle speed. For example, the *word-length effect*—that longer words tend to be fixated more often and for longer (Pollatsek et al., 2008; Rayner & McConkie, 1976)—provides an index of visual and/or early lexical processing. Likewise, the *word-frequency effect* is another common indicator of lexical processing, which refers to the finding that words of low frequency tend to receive more and longer fixations than words of high frequency (Inhoff & Rayner, 1986; Rayner et al., 2004), because it takes longer to retrieve information about uncommon words than common words from memory (Reichle, 2021). Finally, the *wrap-up effect* refers to the finding that, with word frequency and word length being equal, words at the end of a sentence or clause are likely to be recipients of longer fixations compared to words in other locations. Although the wrap-up process is also affected by factors unrelated to high-level integration (e.g., punctuation marks and intonation; see, e.g., Hirotani et al., 2006; Stowe et al., 2018), the wrap-up effect has been traditionally used as an indicator of post-lexical processing wherein readers are converting the linguistic representation of the sentence or clause into a propositional representation for the construction of the situational model of a text (see, e.g., Rayner et al., 2000; Tiffin-Richards & Schroeder, 2018). Liao et al. (2021) found that increasing subtitle speed resulted in attenuated word-frequency effects

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

(indicative of shallower lexical processing) and wrap-up effects (reflecting the incomplete post-lexical integration of the sentences). This suggests that, as subtitle speed increased, readers may have employed some type of text-skimming strategy to rapidly extract the gist of the text, and consequently, that lexical and post-lexical processing of the subtitles were attenuated.

Although Liao et al.'s result that presenting video content with subtitles yields better comprehension than presenting subtitles only is consistent with the prediction of the cognitive theory of multimedia learning, Mayer's (2014) theory does not specify or predict how the processing of one source of information (e.g., the video) might affect the processing of other sources of information (e.g., the subtitle). To provide a more detailed account of how the co-referencing across two visual sources (i.e., background video and subtitle) can facilitate comprehension, and how the presence of video content might influence the reading of subtitles, Liao et al. (2021) proposed a preliminary multimodal integrated-language framework. This theoretical framework extends *E-Z Reader*, a computational model that provides a high-level description of how the cognitive systems responsible for visual processing, attention, language, and oculomotor control are coordinated to support skilled reading (Reichle et al., 2006, 2012). It is also broadly compatible with the dual-coding theory proposed by Paivio (1986, 2007).

An adapted schematic diagram of the framework is presented in Figure 1. (Note that components outside the dashed-line box represent the original framework by Liao et al., 2021, while those inside the dashed-line box illustrate the new additions to the auditory channel in the initial framework. To facilitate our introduction of the framework here, we first focus on the original framework.) As shown, comprehension in multimodal reading is a process of constructing a situational model or mental representation for the integrated information coming from two different processing systems—the auditory system and the visual system. Different systems have sub-systems that are subject to different processing limitations, as indicated by the shading of the boxes. For example, the white boxes correspond to those processes that can only operate upon or actively maintain one representation at any given time (i.e., serial processing). Thus, according to this framework, only one printed word or visual object can be identified via the attention-binding mechanism at any given time because attention must be allocated in a strictly serial manner (Reichle et al., 2009). The gray boxes, however, correspond to those processes that can operate upon or concurrently maintain multiple representations (i.e., parallel processing). The locations of four or five visual objects, for example, can be simultaneously indexed via a pre-attentive stage of visual processing (Pylyshyn, 2004; Pylyshyn & Storm, 1988). This limited parallelism allows the viewer of a video to sidestep the "bottleneck" of attention that results from the serial allocation of attention, thereby allowing them to read the subtitles while, for example, simultaneously monitoring objects in the video in peripheral vision.

Another critical characteristic of this framework is that, while each of the depicted processes performs a specific function, the processes also "communicate" in specific ways. This effectively means that, in multimodal reading situations, the non-textual sources of information that contain redundant (i.e., overlapping or similar) content with the written text can provide additional support for comprehension, thus modulating the reader's need or propensity to read the text. For example,

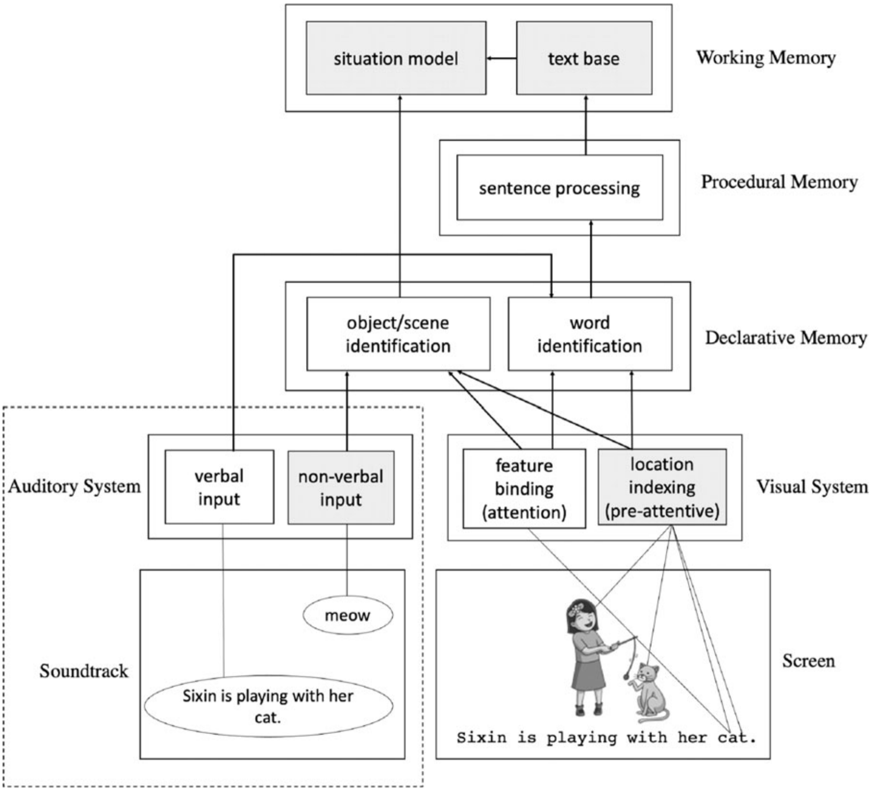


Figure 1. Schematic diagram of the new multimodal integrated-language framework. Components outside the dashed-line box are the same as those in the original framework by Liao et al., 2021, while those inside the dashed-line box are new elements added to the auditory channel. The gray boxes correspond to processes that maintain/operate currently upon multiple representations (i.e., parallel processing), whereas the white boxes correspond to processes that maintain/operate upon only one representation at a time (i.e., serial processing).

unlike the normal reading conditions wherein text comprehension is expected to be severely impaired when word identification (based on the written text only) is inaccurate or slow, the negative ramifications of poor word identification in multimodal reading situations could be compensated for—at least to some extent—by the information available from the other sources of information (e.g., the non-textual visual elements). This thus provides an explanation for Liao et al.'s (2021) finding that participants' comprehension enhanced even when they spent less time reading the subtitles as subtitle speed increased.

Based on the multimodal integrated-language framework, Reichle et al. (2021) used the E-Z Reader model to simulate the possible strategies adopted by Liao et al.'s (2021) participants to compensate for the demands caused by faster subtitle speeds. Using sentences from the Schilling et al. (1998) corpus, they systematically manipulated the model's parameters that control saccadic programming (e.g., fewer refixations, more accurate saccades, and longer preferred saccades) and lexical

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

and/or more complex processing (e.g., imposition of lexical processing deadline and skipping of short words) to determine the processing strategy that would reproduce the qualitative patterns as observed in Liao et al.'s (2021) study—that is, fewer and shorter fixations, as well as longer saccades as subtitle speed increased from 12 to 28 cps. Simulation results showed that promising compensatory strategies include imposing a lexical-processing deadline, a simple word-targeting strategy of skipping short words, or using the video content to enhance the word predictability for difficult (low-frequency) words. These results indicate that, even in the demanding multimodal reading situation (i.e., very fast subtitle speed with video concurrently presented on the screen), readers could still employ some type of strategy to maintain a desirable level of comprehension.

However, it is important to note that, although the multimodal integrated-language framework by Liao et al. (2021) provides a feasible account of how non-text visual elements might support reading comprehension, and Reichle et al.'s (2021) simulations provided some preliminary evidence for its efficacy, it does not specify how verbal and non-verbal information from the auditory system might be integrated with information from the visual system (e.g., the video and subtitles) to support some overall level of video comprehension. Because empirical evidence in this regard is also lacking in Liao et al.'s (2021) study (because their experiment was done without audio), the present study thus provides an opportunity to extend and further evaluate the multimodal integrated-language framework.

Eye-tracking evidence from previous research

Despite the challenges in developing a full account of how perception and cognition are coordinated to support subtitle reading, a growing body of research has provided some clues by examining how background sounds (e.g., music, noise, and speech) affect the reading of static text. These studies generally report a negative effect of the sounds on reading comprehension, with background language being particularly disruptive (see Vasilev et al., 2018, for a review). For example, reading with lyrical music or intelligible speech causes longer sentence-reading times and an increased number of fixations and regressions than reading in silence (Cauchard et al., 2012; Vasilev et al., 2019; Yan et al., 2018; Zhang et al., 2018). In addition, intelligible or meaningful speech generated more disruptions on various eye-movement measures compared with unintelligible speech (e.g., speech in an unknown language: Vasilev et al., 2019; scrambled speech, Yan et al., 2018), suggesting that semantic processing of the auditory content is the primary cause of the auditory disruption effect (Marsh et al., 2008, 2009; Martin et al., 1988).

The study by Hyönä and Ekholm (2016) is most relevant to the present study because in one of their experiments (Experiment 2), they manipulated the similarity of semantic content of the background speech. Specifically, participants were instructed to read texts in three background speech conditions: (1) silence; (2) scrambled speech of the to-be-read text (by randomizing the order of words of the to-be-read text so that the speech is semantically related to the written text); and (3) scrambled speech of an unrelated text (by randomizing the order of words of a text that is unrelated to the to-be-read text). It was found that scrambled speech

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

yielded longer first-pass fixations times compared to the silent condition, but the two types of scrambled speech did not differ in the size of the disruption effect, indicating that the disruption of scrambled speech is not caused by the similarity of semantic content but rather by the fact that both sources of information (the written and spoken text) require the same resource for processing the meaning of words. However, because the speeches used by Hyönä and Ekholm (2016) were anomalous both syntactically and semantically, it remains unclear whether meaningful speech (i.e., semantically and syntactically accurate) with similar or identical semantic content (i.e., semantically relevant) to the written text would result in the disruption effect or not. Our experiment will provide an opportunity to address this question.

Furthermore, it is still inconclusive precisely how the reading process is disrupted by background speech. For example, Yan et al. (2018) found that effects of the frequency of their manipulated target words were only observed in later fixation measures (i.e., gaze duration and total-reading time) but not in first-fixation duration in the presence of background speech (meaningful or meaningless), which was in contrast to silent reading, where the word-frequency effects were observed in both early and late fixation measures. This led Yan et al. to conclude that background noise disrupts the early processing of words during reading. However, Vasilev et al. (2019) observed word-frequency effects in both first- and second-pass eye-movement measures regardless of whether the background sounds were intelligible, unintelligible, or absent. Critically, they found that intelligible speech produced more rereading fixations and regressions as compared to silent reading or reading with unintelligible speech. These findings collectively suggest that intelligible speech does not disrupt the early lexical stage of word identification, but instead interferes with the post-lexical stage of (linguistic) processing (e.g., the integration of words into their sentence contexts).

Although the above studies provide important insights into the influence of audio information on reading, it should be noted that the auditory stimuli used in these studies were mostly scrambled (i.e., meaningless) speeches or speeches that are meaningful but irrelevant to the text being read. As such, the semantic processing of the auditory information likely competes for the cognitive resources that are also involved in reading the text, thereby causing disruption to the reading process. Such situations differ from the normal situation of reading subtitles, where the auditory information is often highly relevant, if not identical, to the content of the text being read.

In contrast to what has been found in the reading of static text, studies on subtitle reading show that people actually spend less time reading subtitles when the audio is present than absent, or in a known than unknown language, indicating the text-relevant auditory information can support, instead of interfering with, subtitle reading. For example, Ross and Kowler (2013) found that viewers spent less time reading subtitles when audio was present compared to the reading-only condition (without audio), due to the skipping of subtitles in the audio-present condition (see also, d'Ydewalle et al., 1991). Similar findings were also reported by Szarkowska and Gerber-Morón (2018) who examined eye movements in conditions with audio in a known versus unknown language. They found that participants spent less time reading subtitles when the audio was in their known (e.g., native or highly proficient second) language than an unknown language. Taken together, these studies provide

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

evidence for the auditory support for subtitle reading—that is, for the hypothesis that viewers rely less on the reading of subtitles for comprehension when relevant verbal information is available from the auditory input.

However, it remains unclear *how* the auditory context would support subtitle reading. For example, does auditory processing support the lexical and/or post-lexical processing of subtitles? Our lack of understanding here is largely due to the fact that previous studies on subtitle reading have mostly reported only global eye-movement measures on the entire subtitle (see Table 1 for examples) and thus provide no indication of precisely how the presence versus absence of concurrent auditory information actually affects the processing of the subtitles. A global increase in sentence-reading times, for example, might reflect an overall slowing in lexical processing (e.g., increased fixation durations), more difficult post-lexical processing (e.g., more inter-word regressions), or both.

Another limitation in the existing research on subtitle reading is that the studies reviewed above all presented subtitles in participants' native language. For studies that used non-native language subtitles (see Table 1), subtitles were in a language that participants had no or little knowledge of (see, e.g., Bisson et al., 2014; d'Ydewalle & De Bruycker, 2007). One exception is a recent study by Ragni (2020), which examined second-language subtitles with native-language audio. However, as Ragni's study did not provide a control condition (e.g., the no-audio condition), it is difficult to determine the influence of audio on the reading of second-language subtitles, although such experimental design serves its purpose of investigating the impact of translation strategies on the processing of second-language subtitles. Overall, there is limited empirical evidence for how the auditory input may influence subtitle reading when subtitles are in readers' second language. Answers to this question will have significant practical implications for the application of subtitling in real-world activities given the widespread use of second-language subtitles in educational videos as a tool to assist learning for international students.

Given this brief overview of the work that has been done to examine the influence of auditory information on reading, our study aims at filling in the research gaps identified in two closely related disciplines—research on reading static text, where the influence of semantically relevant audio on reading is underexplored, and research on subtitle reading, where the influence of semantically relevant audio on reading has been impeded by methodological limitations. To this end, we made the first attempt to understand how semantically relevant auditory information will influence reading in the context of watching video with subtitles, using both global and local (i.e., word-based) eye-movement measures. We did this by manipulating three types of auditory information that were expected to modulate the necessity for reading (and understanding) English/L2 subtitles: (1) Chinese/L1 audio (i.e., interlingual subtitles with audio in a different language), which obviated the need to read the English subtitles; (2) English/L2 audio (i.e., intralingual subtitles with audio in the same language), which may facilitate the reading of the English subtitles; and (3) no audio, which necessitated subtitle reading if participants need to fully understand the video content. We also manipulated the presence versus absence of the concurrent video content to examine how the interaction between auditory processing and

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

Table 1. Summary of major eye-tracking studies on the influence of audio on subtitle reading

Study	Subtitle and Audio	Participants	Eye-tracking measures[4]
d'Ydewalle et al. (1987)	Dutch subtitle with or without German audio	Dutch speakers with little or good knowledge of German	- Percentage dwell time[5] - Total-reading times
d'Ydewalle et al. (1991)	**Exp. 1:** English subtitle with or without English audio	English speakers (unfamiliar with subtitle reading)	- First-fixation latency - Percentage dwell time (subtitle and video)
	Exp. 2: Dutch subtitle with or without Dutch audio	Dutch speakers (familiar with subtitle reading)	
d'Ydewalle & De Bruycker (2007)	1) Dutch subtitle with Swedish audio; 2) Swedish subtitle with Dutch audio	Dutch speakers with no knowledge of Swedish	- First-fixation latency - Mean fixation durations - Mean for/backward saccade amplitudes - Number of subtitle-video crossovers - Percentage dwell time - Regression rate - Skipping rate - Word fixating rate
Ross & Kowler (2013) (Exp.1)	1) English audio only; 2) No audio and no subtitle; 3) English subtitle with English audio; 4) English subtitle without audio	English speakers	- Duration of crossover saccades - First-fixation latency - Landing-position distribution (video) - Mean fixation durations (subtitleand video) - Mean forward saccade amplitudes - Number of video-to-subtitle crossovers - Percentage dwell time (subtitle and video) - Proportion of subtitle-video

(*Continued*)

Table 1. (*Continued*)

Study	Subtitle and Audio	Participants	Eye-tracking measures[4]
			crossovers - Regression rate
Bisson et al. (2014)	1) English subtitle + Dutch audio; 2) Dutch subtitle + English audio; 3) Dutch subtitle + Dutch audio; 4) Dutch audio	English speakers with no Dutch knowledge	- Mean fixation durations (subtitle/video) - Number and proportion of consecutive fixations - Skipping rate - Total number of fixations (subtitle and video) - Total viewing times (subtitle and video)
Lång (2016) (Exp. 2)	Finnish subtitle with Russian audio	Finnish speakers with no Russian knowledge; Russian speakers with no Finnish knowledge	- Mean fixation durations (subtitle/video) - Number of subtitle-video crossovers - Total number of fixations (subtitle and video) - Total viewing times (subtitle and video)
Szarkowska & Gerber-Morón (2018)	**Exp. 1**: Native-language subtitle (English/Polish/Spanish) with Hungarian audio	English/Polish/Spanish speakers with no Hungarian knowledge	- Mean fixation durations - Number of revisits of entire subtitles - Percentage dwell time - Total number of fixations - Total-reading times
	Exp. 2: Native-language subtitle (English/Polish/Spanish) with English audio	English/Polish/Spanish speakers with high English proficiency	
Ragni (2020)	Italian subtitle with English audio	English learners of Italian	- Mean fixation duration - Total number of fixations

subtitle reading might be modulated by the overall visual-processing demands and/or a (partially) redundant source of information.

Based on the multimodal integrated-language framework (Liao et al., 2021), which postulates that viewers can perform limited parallel processing (e.g., tracking a smaller number of objects in the background video using peripheral vision while reading subtitles) and can combine different sources of information in response to varying task demands to optimize comprehension, it is hypothesized that, on a global level, participants would rely less on subtitles when audio provides an additional source for identical or similar verbal information. Therefore, we expected fewer, shorter fixations, longer saccades as well as more skipping of subtitles with Chinese or English audio compared to without audio (Hypothesis 1). On a local level, as native-language audio presumably eliminated the need for subtitles, it is hypothesized that lexical and post-lexical processing of subtitles would be attenuated with Chinese audio than without audio (Hypothesis 2). However, participants might use the second-language audio to support the reading of second-language subtitles (or vice versa), thereby allowing them to engage in deeper lexical and post-lexical processing of subtitles with English audio than without audio (Hypothesis 3). Finally, the effects of audio on subtitle reading are likely to be more evident in the presence of concurrent video content (as opposed to the absence of video) which provides an additional source of overlapping information for comprehension (Hypothesis 4).

Method

Participants

Thirty-four Chinese native speakers who were also advanced speakers of English (scoring 6 or 7 in the Reading and Listening bands in the *International English Language Testing System*, or *IELTS*) were recruited as participants (26 females). Their average age was 25.8 years (*SD* = 3.98, *range* = 20–38). All participants reported normal or corrected-to-normal vision. Ethics approval was obtained from Macquarie University (Reference No: 5201830023375). Participants were awarded cash or course credit in accordance with ethical requirements.

Design

The experiment was a 3 (audio condition: Chinese/L1 audio, English/L2 audio, no audio) × 2 (video condition: present vs. absent) within-subject design, resulting in six experimental conditions. All conditions were counterbalanced via a Latin-square design to ensure that each participant watched each video in a given condition once. A Latin-square design was also used to assign each video to each of the six conditions equally often across participants. Videos were presented to participants in a random order.

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

Table 2. Characteristics of video clips and subtitles

Video clip	Duration (mins)	Number of subtitles	Number of words	Average characters per subtitle	Average duration per subtitle (seconds)	Average subtitle speed (cps)	Flesch Reading Ease
From Pole to Pole	8.58	86	596	35	2.71	12	91.01
Mountains	8.16	85	537	36	2.60	13	79.89
Deserts	8.07	78	517	36	2.71	13	79.38
Great Plains	9.30	88	560	35	2.80	12	80.16
Shallow Seas	10.14	89	563	34	2.73	12	78.59
Seasonal Forest	8.46	91	640	36	2.90	12	83.80

Note. Subtitle speed is measured by the number of characters presented per second (cps).

Materials

Stimuli

Six video clips (each of 8–10 min) were selected from six episodes of the BBC documentary series *Planet Earth* (Fothergill, 2006) as stimuli. The videos had no on-screen speakers, which makes it possible to change the language in the soundtrack without causing problems with lip synchronization. All video clips were self-contained and comparable in terms of the narrative structure, pace of spoken dialogue, and visual complexity. The linguistic complexity (i.e., reading ease) of the subtitles from the different videos was compared using *Coh-Metrix* (http://tool.cohmetrix.com/), a computational tool for readability testing (Graesser et al., 2014) (see Table 2).

English subtitles for all video clips were generated as verbatim transcripts of the original English audio using the *Aegisub* subtitle-editing software[1] and guidelines listed in Table 3. Chinese audio, which was a direct translation of the original English audio, was extracted from the same documentary series broadcast by the Central Broadcasting Television (CCTV), the predominant state television broadcaster in Mainland China. Like the English audio, the Chinese audio was synchronized with the onset of the subtitle, and the semantic meaning of each subtitle was equivalent to the meaning of its corresponding Chinese audio for most subtitles[2] (see Table 4 for examples of subtitles and auditory transcription). Subtitles that did not have the same semantic meanings with the Chinese audio were excluded in eye-movement analyses for all experimental conditions (13% data loss in global analyses and 11% in local analyses). In this way, the influence of the semantic discrepancy between the two information sources (i.e., the subtitle and the audio) was minimized. Apart from the languages being used (i.e., English vs. Chinese), the two audio conditions were exactly the same with respect to other non-verbal content (e.g., background noises).

Table 3. Guidelines for the generation of subtitles in the current study

Parameters	Guidelines
Onset time	Subtitles came on when the speech started and went off when the speech went off
Number of lines	One line
Presentation speed	No more than 20 cps ($M = 12.35$, $SD = 1.54$; range: 6–18 cps)
Number of characters per line	No more than 55 characters
Sentence break	One sentence spread over no more three subtitles. Sentences were broken at logical points so that each subtitle formed a comprehensible segment

Table 4. Examples of English subtitles and Chinese audio used in the study (back translation of the Chinese audio in square brackets)

English subtitle	Chinese audio [back translation]
It's March and light returns to the high Arctic	三月来了,阳光回到了北极 [March comes and light returns to the Arctic]
Sweeping away four months of darkness	驱走了长达4个月的黑暗。 [dispelling four months' darkness]
A polar bear stirs	北极熊出来了。 [polar bear comes out]
She has been in her den the whole winter	它整个冬天都在洞里度过。 [it spent whole winter in the den]
Her emergence marks the beginning of spring	北极熊走出洞穴意味着春天来了。 [Polar bear coming out of the den means that spring has arrived.]

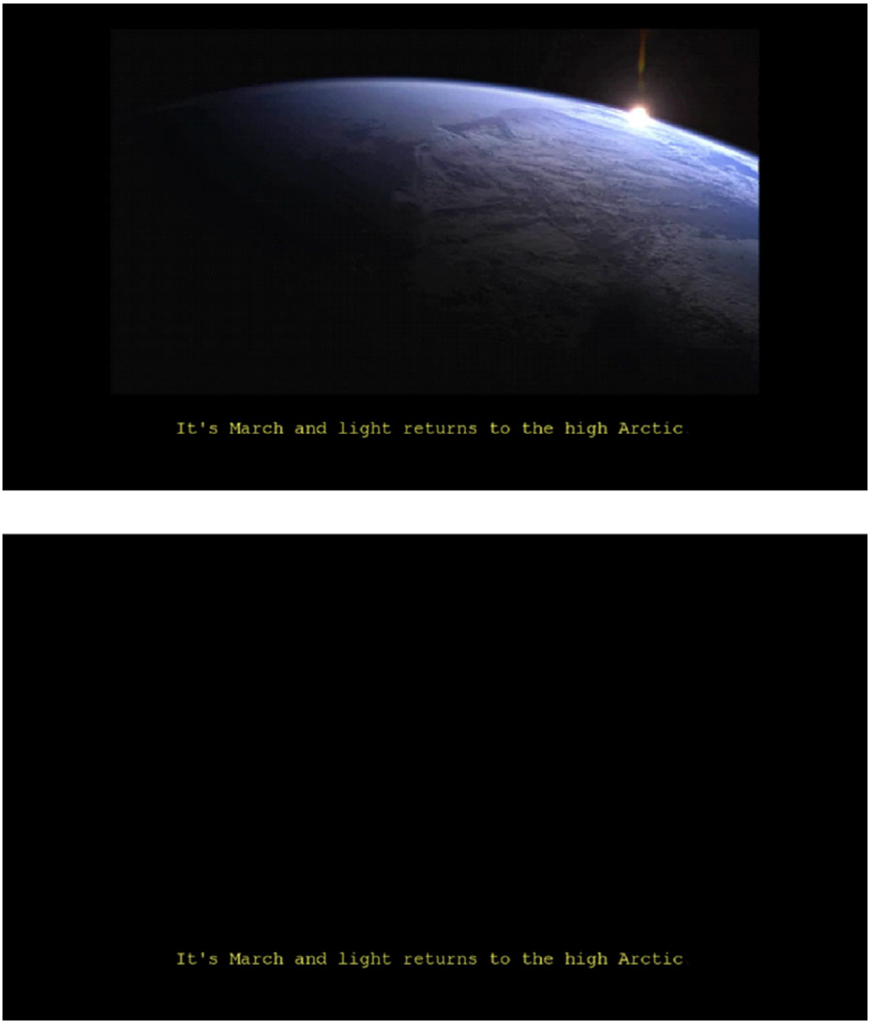
Comprehension tests

Eight three-alternative-choice questions derived from the subtitles only were used to evaluate participants' comprehension of each video. All questions were presented in bilingual scripts to preclude the confounding influence of language on comprehension results.

Apparatus and procedure

Participants' eye movements during subtitle reading and video viewing were recorded using an EyeLink 1000+ (SR Research Ltd., Canada) eye tracker with a sampling rate of 2,000 Hz. Stimuli were displayed on a BenQZowie XL2540 screen with a refresh rate of 240 Hz and a screen resolution of 1,920 × 1,080 pixels. Videos were presented with a resolution of 1,280 × 720 pixels and a presentation rate of 30 frames per second at the center of the screen, and subtitles were presented below the

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press



It's March and light returns to the high Arctic

It's March and light returns to the high Arctic

Figure 2. Screenshots of stimuli in the video-present condition (at the top) and video-absent condition (at the bottom).

video in mono-spaced Courier New font (30-point; RGB color: 255, 255, 102) (see Figure 2). Each subtitle was displayed on the screen one line at a time in synchrony with the audio, which was played with the volume level of 75–80 dBA via two external speakers that were placed on each side of the computer monitor in a sound-proof laboratory. A chin-and-forehead rest was used to minimize head movements. Participants were seated 95 cm away from the monitor, which produced a ~0.4° visual angle for each letter on the screen. Viewing was binocular, but only the right eye was tracked.

Participants were tested individually in a sound-proof and sufficiently illuminated laboratory. Prior to the eye-tracking experiment, participants were given a

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

participant consent form and a brief verbal task instruction. Participants were not instructed to pay specific attention to subtitles. To ensure tracking accuracy, a nine-point calibration and validation procedure was performed prior to watching each video (the maximum allowance for the calibration error was 0.5°). Comprehension questions were presented at the center of the screen one by one after each video. Participants were given a 2-min break after watching each video to avoid fatigue. The whole experiment lasted approximately 2.5 hr.

Analyses

Three participants' data were removed prior to the analyses due to tracking loss in some of the videos, resulting in 31 participants' data being used for eye-movement analyses. Fixations shorter than 60 ms or longer than 800 ms were also excluded (8.68%) from the analyses.

Several global eye-movement measures of which the analyses were based on the entire subtitle region were reported: (1) average fixation durations, (2) total number of fixations, (3) progressive saccade length (i.e., average length of rightward saccades in degrees), and (4) percentage of skipped subtitles (i.e., percentage of entire subtitles that are not fixated). These global measures provide us with a general pattern of how the subtitle reading is modulated by participants' needs for the subtitles in different audio conditions, and the potential strategy adapted by participants. For example, in Liao et al.'s (2021) study, the combination of shorter and fewer fixations, and longer saccade length is indicative of a skimming strategy.

In order to gain more insights into the lexical and post-lexical processing of subtitles, we also examined the word-frequency, word-length, and wrap-up effects using two word-based fixation measures that are commonly used in reading research: *gaze durations* (i.e., the sum of all fixations on a word prior to the eyes exiting the word) and *total-reading times* (i.e., the sum of all fixations on a word). Examination of these two measures allows us to probe into the influence of audio on different stages of linguistic processing, with gaze durations reflecting early stages and total-reading times reflecting relatively late stages of linguistic processing (including regressions back to words; Rayner, 1998, 2009).

Data were analyzed via generalized/linear mixed models (G/LMMs) using the *lme4* package (version 1.1-23) in *R* (Version 3.6.3); *p* values were computed via the *lmerTest* package (Version 3.1-2, Kuznetsova et al., 2017). Fixations were log-transformed in all analyses to meet the data assumption of the LMMs analyses. Main effects of experimental factors were extracted using sliding difference contrasts via the *contr.sdif* function, which compares consecutive factor levels of each variable (i.e., Chinese audio vs. English audio and English audio vs. no audio, for the audio condition comparison). The *emmeans* package (version 1.47) was used to compute and extract the estimated means between the Chinese-audio and no-audio conditions, and the simple effects.

When fitting a model, we first started with a maximal random-effect structure that included all experimental factors and their interactions as fixed effects, subject and subtitle item (or word item in the word-based analyses) as random intercepts, as well as random slopes for the fixed effects across subjects and subtitles/words (Barr et al., 2013). Because all video clips were selected from the same documentary series

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

and thus homogeneous (with linguistic complexity, video duration, and genre being comparable), a single random-effect variable was coded for each combination of video and subtitles in our global analyses, and for each combination of video, subtitle, and word in our word-based analyses. Insignificant random-effects components were progressively removed to generate a parsimonious model (Bates et al., 2015; see Appendix 1 for a summary of final models used in data analyses). Both word frequency and length were entered into the models as scaled and centered continuous variables; the word frequency is based on the *Zipf* scale, or log10 (frequency per billion words), from the *SUBTLEX-UK* word-frequency corpus (van Heuven et al., 2014). For word-frequency/length effect analyses, the first and final words in each subtitle were excluded to avoid any potential confounds due to the sudden appearance or disappearance of the subtitle.

Because real-world (i.e., commercially available) videos were used as stimuli in the present study, the wrap-up effects were examined by comparing words in two locations: line ending versus non-ending, which includes all words from the second to the penultimate in the subtitles (cf., examine the wrap-up effect with experimental manipulation; Warren et al., 2009)[3]. The average zipf frequencies for the ending versus non-ending words were 4.5 versus 5.5, and their average lengths were 6.1 versus 4.9 letters, respectively.

Finally, to control for potential type I error associated with the use of multiple eye-tracking measures, Bonferroni correction was applied as a remedy by dividing the 0.05 alpha threshold by the number of dependent measures used to examine a given effect (von der Malsburg & Angele, 2017). This yielded an alpha level of 0.013 for our global analyses (four measures used) and alpha of 0.025 for the two word-based analyses (two measures used).

Results

Comprehension

Mean comprehension accuracy across six conditions for each participant was above chance level (0.33), ranging from 0.60 to 0.85 ($M = 0.72$, $SD = 0.45$) across participants. Figure 3 shows that participants had higher comprehension accuracy with concurrent video content ($z = 3.10$, $p < 0.05$). No main effect of audio condition was observed (all $|z|$s < 1.57, ps > 0.05), which indicates that participants obtained similar levels of comprehension irrespective of whether or not there was semantically relevant audio.

Global analyses of eye movements

Mean fixation durations

As Tables 5 and 6 and Figure 4A show, participants made shorter fixations on subtitles when the video was present. There was no main effect of the audio condition, but interaction between video and audio (Chinese vs. no audio) was observed. Pairwise contrasts of the Video × Audio (Chinese vs. no audio) interaction revealed that fixations were shorter with Chinese audio than without audio, but only when video was present (video absent: $t = 0.75$, $p = 0.45$; video present: $t = -3.35$,

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

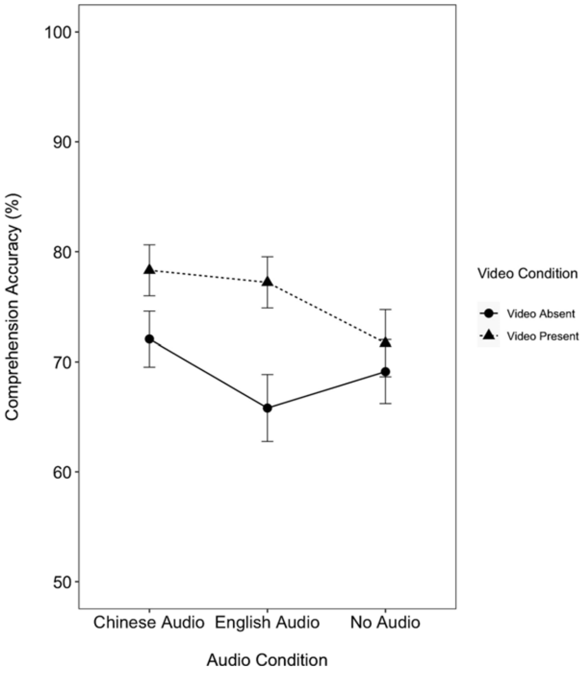


Figure 3. Comprehension accuracy as a function of video and audio conditions. Error bars represent the standard errors of the means.

Table 5. Mean and standard deviation of global eye-movement measures in the subtitle region.

Video	Audio	Average fixation durations (ms)	Total number of fixations	Progressive saccade length (in degrees)	Percentage of skipped subtitles
Absent	CA	235 (36)	7.09 (2.19)	2.67 (0.31)	0.06 (0.10)
Absent	EA	227 (28)	7.97 (1.79)	2.49 (0.28)	0.04 (0.11)
Absent	NA	227 (23)	9.03 (1.56)	2.43 (0.32)	0.02 (0.07)
Present	CA	207 (32)	5.15 (1.86)	2.79 (0.33)	0.15 (0.14)
Present	EA	209 (27)	6.71 (1.25)	2.63 (0.38)	0.02 (0.04)
Present	NA	216 (26)	8.06 (1.25)	2.53 (0.34)	0.01 (0.02)

Note. CA, Chinese audio; EA, English audio; NA, no audio.

$p < 0.001$). Finally, there was no significant difference between English audio and no audio.

Total number of fixations

As Tables 5 and 6 and Figure 4B show, there were fewer fixations in the subtitle region with video presentation than without. Participants also made fewer fixations

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

Table 6. The LMMs results for the global analyses in the subtitle region

Measures	Contrasts	*b*	*SE*	*t/z*	*p*
Average **f**ixation **d**urations	Intercept	5.37	0.02	259.63	< **.001**
	Video (present–absent)	−0.09	0.01	−5.85	< **.001**
	Audio (EA–CA)	0.00	0.01	−0.08	0.94
	Audio (NA–EA)	0.02	0.01	1.78	.09
	Audio (CA–NA)	−0.02	0.01	−1.32	0.19
	Video × Audio (EA–CA)	0.04	0.02	2.38	.02
	Video × Audio (NA–EA)	0.03	0.02	1.49	0.15
	Video × Audio (CA–NA)	0.07	0.03	2.63	**.01**
Total **n**umber of **f**ixations	Intercept	7.33	0.23	31.31	< **.001**
	Video (present–absent)	−1.35	0.27	−4.95	< **.001**
	Audio (EA–CA)	1.24	0.20	6.09	< **.001**
	Audio (NA–EA)	1.16	0.20	5.68	< **.001**
	Audio (CA–NA)	−2.40	0.25	−9.53	<**.0001**
	Video × Audio (EA–CA)	0.69	0.37	1.84	0.08
	Video × Audio (NA–EA)	0.36	0.32	1.11	0.28
	Video × Audio (CA–NA)	1.04	0.39	2.69	**.01**
Progressive **s**accade **l**ength	Intercept	0.81	0.02	42.88	< **.001**
	Video (present–absent)	0.05	0.01	4.25	< **.001**
	Audio (EA–CA)	−0.05	0.01	−3.98	< **.001**
	Audio (NA–EA)	−0.02	0.01	−1.61	0.12
	Audio (CA–NA)	0.07	0.01	5.73	<**.0001**
	Video × Audio (EA–CA)	0.02	0.01	1.75	.08
	Video × Audio (NA–EA)	−0.02	0.01	−1.43	0.15
	Video × Audio (CA–NA)	0.01	0.01	0.49	0.63
Percentage of **s**kipped **s**ubtitles	Intercept	−5.18	0.37	−14.05	<**.0001**
	Video (present–absent)	1.50	0.54	2.79	**0.005**
	Audio (EA–CA)	−2.00	0.38	−5.22	<**.0001**
	Audio (NA–EA)	−0.70	0.56	−1.26	0.21
	Audio (CA–NA)	2.70	0.40	6.73	<**.0001**
	Video × Audio (EA–CA)	−2.25	0.31	−7.14	<**.0001**
	Video × Audio (NA–EA)	1.69	0.47	3.59	<**.0001**
	Video × Audio (CA–NA)	−0.55	0.39	−1.42	0.16

Note. Bold font indicates $p < .013$. CA, Chinese audio; EA, English audio; NA, no audio.

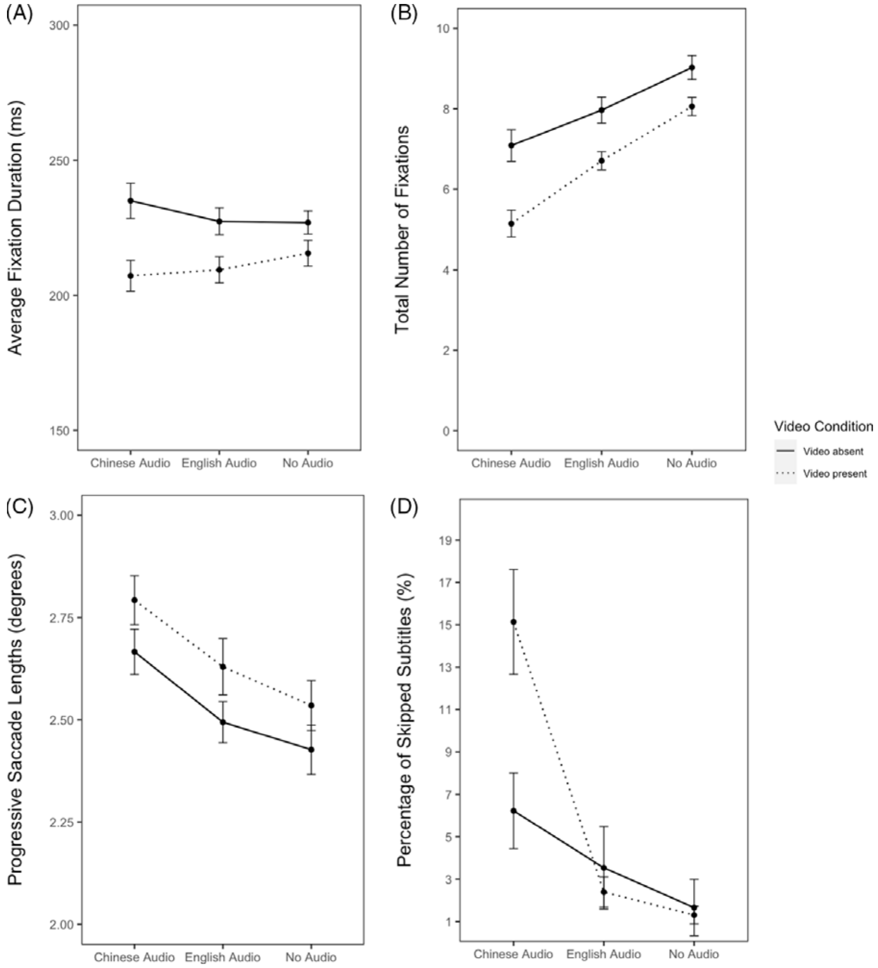


Figure 4. Mean descriptive statistics for global eye-movement measures in the subtitled region. Error bars represent the standard errors of the means.

on the subtitles as the need or propensity to read them decreased (i.e., moving from the no- to English- to Chinese-audio conditions). The difference between Chinese audio and no audio was more pronounced with video presentation, as revealed by the Video × Audio (Chinese vs. no audio) interaction (video absent: $b = -1.88$, $t = -5.36$, $p < 0.0001$; video present: $b = -2.92$, $t = -10.36$, $p < 0.0001$).

Progressive saccade length

As shown in Tables 5 and 6 and Figure 4C, participants made longer progressive saccades in the subtitle region when video was present than absent. Saccades were also longer with Chinese audio than with English audio or without audio.

Percentage of skipped subtitles

As shown in Tables 5 and 6 and Figure 4D, more subtitles were skipped when concurrent video was present than absent. Participants skipped more subtitles when the need or propensity to read them decreased (i.e., with Chinese audio compared to with English audio or without audio). But the difference between Chinese audio and English audio was only significant when video was present (video absent: $t = 2.06$, $p = 0.04$; video present: $t = 7.76$, $p < 0.001$), as revealed by the Video × Audio (Chinese vs. English) interaction. The contrast between English audio and no audio also interacted with video condition. Pairwise analyses of the Video × Audio (English vs. no audio) interaction showed that the difference between English audio and no audio was only numerical but did not reach significance in any video conditions (video absent: $t = 2.40$, $p = 0.02$; video present: $t = -0.26$, $p = 0.79$).

Local analyses of eye movements

Word-frequency and word-length effects

Gaze durations. As shown in Table 7 and Figures 5A and 6A, main effects of word frequency, word length, video condition, and audio condition (Chinese vs. English) were observed. Gaze durations on words decreased with increasing word frequency, decreasing word length, and concurrent video content. Participants also fixated words in the subtitles shorter when the audio was in Chinese than in English. The absence of Frequency × Audio and Length × Audio interactions indicated the word-frequency and word-length effects were similar across three audio conditions. Finally, the Frequency × Length × Video interaction indicated that the word-frequency effect increased with increasing word length, but only in the condition without video. Similarly, the word-length effect increased as words became infrequent, but only in the absence of video. The three-way interaction between word frequency, word length, and video condition produced patterns that are consistent with those observed in Liao et al. (2021).

Total-reading times. Main effects of word frequency, word length, video, and audio conditions (Chinese vs. no audio; English vs. no audio) were found, as shown in Table 7 and Figures 5B and 6B. Total-reading times on words decreased with increasing word frequency, decreasing word length, concurrent video content, and low subtitle-reading need or propensity (i.e., when either the English or Chinese audio was available). The Length × Video interaction showed that the length effect was reduced with video presentation, which is again consistent with the results of Liao et al. (2021). Audio condition did not interact with word length or frequency, indicating that the word-length and word-frequency effects did not differ across three audio conditions.

Wrap-up effect

Gaze durations. As shown in Table 8 and Figure 7A, main effects of word location (i.e., wrap up), video, and audio conditions were observed. The Location × Video interaction showed that the wrap-up effect was only visible when video was absent (video absent: $t = 6.55$, $p < 0.0001$; video present: $t = 0.93$, $p = 0.35$). The

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

Table 7. The LMMs results for the word-frequency and word-length effects

Measures	Contrasts	*b*	*SE*	*t*	*p*
Gaze durations	(Intercept)	5.49	0.02	262.23	<**.0001**
	Freq.	−0.09	0.01	−12.05	<**.0001**
	Len.	0.10	0.01	12.49	<**.0001**
	Video (present–absent)	−0.06	0.01	−5.40	<**.0001**
	Audio (EA–CA)	0.04	0.01	2.78	**.01**
	Audio (NA–EA)	−0.01	0.01	−0.46	0.64
	Audio (CA–NA)	−0.03	0.02	−1.89	.06
	Freq. × Len.	−0.01	0.00	−3.09	**.002**
	Freq. × Video	0.00	0.01	−0.05	0.96
	Len. × Video	0.01	0.01	1.59	0.11
	Freq. × Audio (EA–CA)	−0.01	0.01	−1.24	0.22
	Freq. × Audio (NA–EA)	0.00	0.01	−0.05	0.96
	Freq. × Audio (CA–NA)	0.01	0.01	1.27	0.20
	Len. × Audio (EA–CA)	0.00	0.01	0.18	0.86
	Len. × Audio (NA–EA)	−0.01	0.01	−0.97	0.33
	Len. × Audio (CA–NA)	0.01	0.01	0.72	0.47
	Video × Audio (EA–CA)	0.05	0.02	2.70	**.007**
	Video × Audio (NA–EA)	0.01	0.02	0.74	0.46
	Video × Audio (CA–NA)	0.06	0.02	3.33	<**.0001**
	Freq. × Len. × Video	0.02	0.01	2.55	**.01**
Total times	(Intercept)	5.78	0.02	253.67	<**.0001**
	Freq.	−0.14	0.01	−14.14	<**.0001**
	Len.	0.16	0.01	18.51	<**.0001**
	Video (present–absent)	−0.19	0.02	−8.89	<**.0001**
	Audio (EA–CA)	−0.01	0.02	−0.43	0.67
	Audio (NA–EA)	0.07	0.01	5.21	<**.0001**
	Audio (CA–NA)	−0.06	0.02	−2.91	**.004**
	Freq. × Len.	−0.01	0.00	−1.56	0.12
	Freq. × Video	0.00	0.01	−0.39	0.70
	Len. × Video	−0.03	0.01	−3.28	**.001**
	Freq. × Audio (EA–CA)	−0.02	0.01	−1.77	.08
	Freq. × Audio (NA–EA)	0.00	0.01	−0.24	0.81
	Freq. × Audio (CA–NA)	0.02	0.01	1.99	.05
	Len. × Audio (EA–CA)	−0.02	0.01	−1.32	0.19
	Len. × Audio (NA–EA)	0.01	0.01	0.48	0.63

(*Continued*)

Table 7. (*Continued*)

Measures	Contrasts	*b*	*SE*	*t*	*p*
	Len. × Audio (CA–NA)	0.01	0.01	0.91	0.37
	Video × Audio (EA–CA)	0.05	0.02	2.88	**.004**
	Video × Audio (NA–EA)	0.05	0.02	2.85	**.004**
	Video × Audio (CA–NA)	0.10	0.02	5.45	<**.0001**
	Freq. × Len. × Video	0.01	0.01	2.16	.03

Note. Bold font indicates $p < 0.025$. CA, Chinese audio; EA, English audio; NA, no audio. Three-way and four-way interactions without significance were not reported for simplicity.

Location × Audio (Chinese vs. English) × Video interaction showed that the Location × Audio (Chinese vs. English) interaction was only significant when video was absent (video absent: $t = -6.88$, $p < 0.0001$; video present: $t = -1.59$, $p = 0.11$). Without video, the wrap-up effect was more pronounced with English audio compared to Chinese audio (Chinese audio: $b = 0.14$, $t = 4.28$, $p < 0.0001$; English audio: $b = 0.28$, $t = 8.73$, $p < 0.001$). The Location × Audio (English vs. no audio) × Video interaction revealed that the Location × Audio (English vs. no audio) interaction was significant in both video conditions but was more pronounced in the absence of video. When video was absent, the wrap-up effect was larger with English audio than without audio (English audio: $b = 0.28$, $t = 8.73$, $p < 0.0001$; no audio: $b = 0.17$, $t = 5.35$, $p < 0.0001$); however, when video was present, the wrap-up effect was only observed without audio but not with English audio (English audio: $b = 0.02$, $t = 0.77$, $p = 0.44$; no audio: $b = 0.07$, $t = 2.11$, $p = 0.03$). Finally, the wrap-up effect was attenuated with Chinese audio than without audio (Chinese audio: $b = 0.07$, $t = 2.14$, $p = 0.03$; no audio: $b = 0.12$, $t = 3.93$, $p < 0.0001$), as revealed by the Location × Audio (Chinese vs. no audio) interaction.

Total-reading times. Main effects of video and audio were observed, but no main effect of word location, as shown in Table 8 and Figure 7B. The Location × Video interaction showed that subtitle-ending words received numerically longer reading times than middle words (a normal wrap-up effect) when video was absent, whereas a reversed trend as observed when video was present, although none of these differences reached significance (video absent: $t = 2.03$, $p = 0.04$; video present: $t = -0.89$, $p = 0.38$). Likewise, the Location × Audio (Chinese vs. no audio) interaction revealed that subtitle-final words were fixated longer than middle words numerically (i.e., the normal wrap-up effect) without audio whereas a reversed pattern was observed for Chinese audio, although the effect of word location was not significant in these two audio conditions (Chinese audio: $t = -0.63$, $p = 0.53$; no audio: $t = 0.51$, $p = 0.61$). Similar to the patterns in gaze durations, Location × Audio (Chinese vs. English; English vs. no audio) × Video interactions were observed. Pairwise contrasts showed that the Location × Audio (Chinese vs. English; English vs. no audio) interactions were only significant when video was absent (video absent: $|t|s > 5.02$, $ps < 0.0001$; video present: $|t|s < 1.67$, $ps > 0.10$). Without video, the wrap-up effect was observed with English

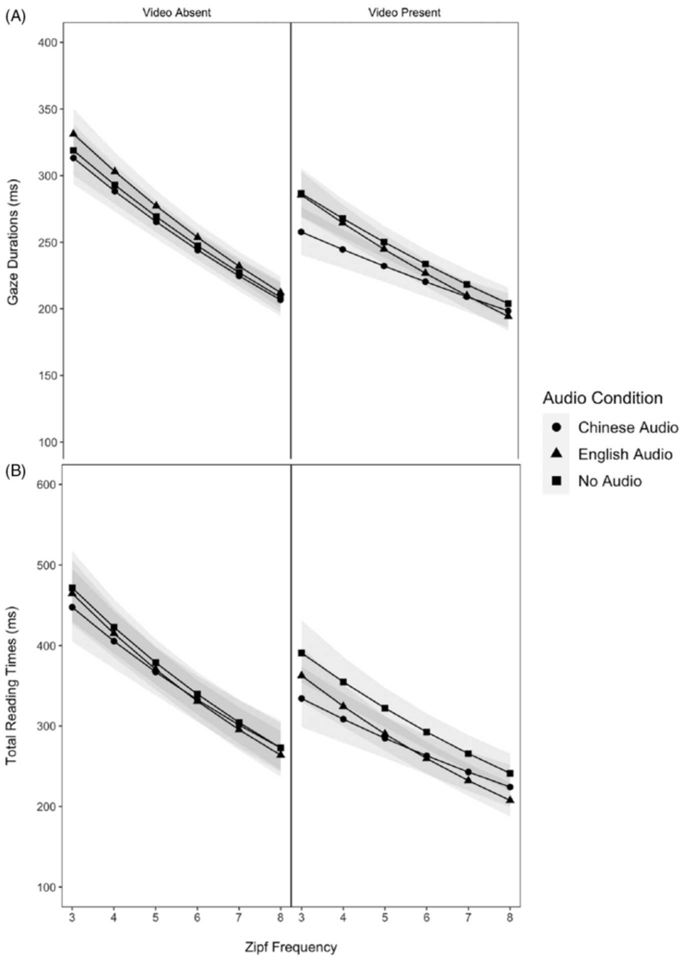


Figure 5. The LMMs-adjusted word-frequency effects as a function of video presence/absence and audio condition. Word frequency is based on the *Zipf* scale extracted from the SUBTLEX-UK word-frequency corpus. Ribbons represent the lower (5%) and upper limits (95%) of confidence intervals for the estimated marginal means.

audio but not with Chinese audio or without audio (Chinese audio: $t = 0.24$, $p = 0.81$; English audio: $t = 4.12$, $p < 0.0001$; no audio: $t = 1.44$, $p = 0.15$).

General discussion

While some has been learned about how semantically irrelevant background speech affects reading, relatively little is known about how reading might be affected by speech that is semantically relevant to the text being read. The answer to this question will have significant practical implications because reading with semantically relevant audio is a common scenario in our daily life, such as reading subtitles when watching videos with soundtrack in a known language. The present study therefore aimed to understand how

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

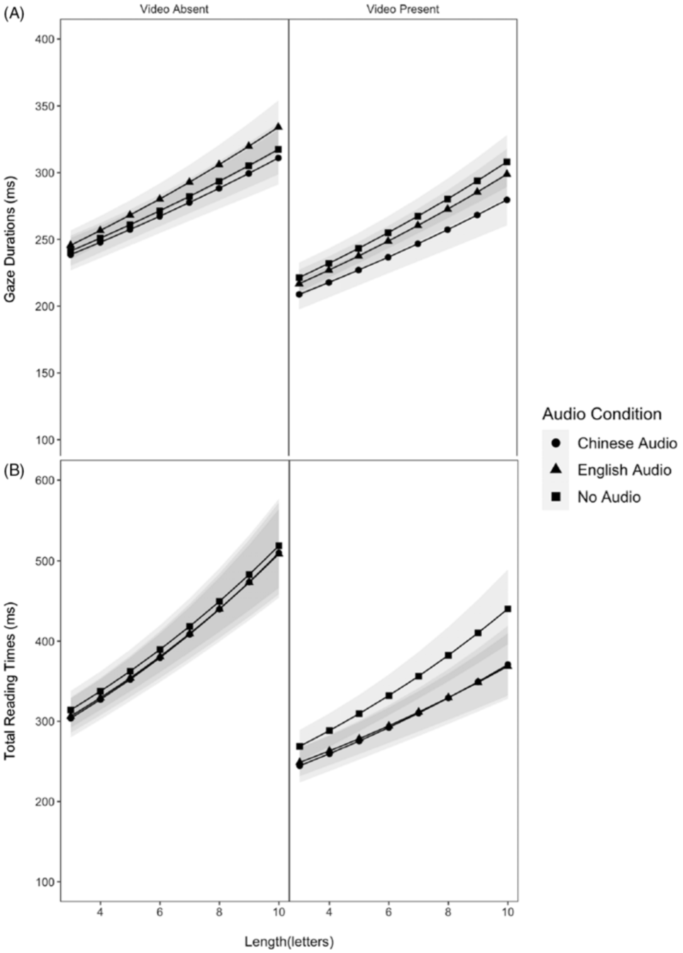


Figure 6. The LMMs-adjusted word-length effects as a function of video presence/absence and audio condition. Ribbons represent the lower (5%) and upper limits (95%) of confidence intervals for the estimated marginal means.

the reading process could be affected by semantically relevant auditory input in the context of reading English/L2 subtitles in video. A 2 (video condition: absence vs. presence) × 3 (audio condition: Chinese/L1 audio vs. English/L2 audio vs. no audio) eye-tracking experiment was conducted in which the manipulation of the audio and video conditions likely modulated the need or propensity to read the subtitles, with lowest propensity occurring with Chinese audio (because the participants were native Chinese speakers) and the highest propensity occurring without audio (because much of the video content could then only be extracted from the English subtitles).

Although there was no evidence from the present study that semantically relevant audio affects reading comprehension, our eye-movement data clearly show that readers adjusted the way they engaged in the reading of subtitles in response to the varying needs to read the subtitles in different audio conditions. Analyses of global

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

Table 8. The LMMs results for wrap-up effects

Measures	Contrasts	*b*	*SE*	*t*	*p*
Gaze durations	(Intercept)	5.53	0.02	230.36	<**.0001**
	Location (middle–end)	−0.11	0.03	−3.83	<**.001**
	Video (present–absent)	−0.15	0.01	−11.93	<**.0001**
	Audio (EA–CA)	0.08	0.01	7.55	<**.0001**
	Audio (NA–EA)	−0.03	0.01	−2.52	**.02**
	Audio (CA–NA)	−0.05	0.01	−4.29	<**.0001**
	Location × Video	0.17	0.01	14.06	<**.0001**
	Location × Audio (EA–CA)	−0.09	0.01	−5.98	<**.0001**
	Location × Audio (NA–EA)	0.03	0.01	2.40	**.02**
	Location × Audio (CA–NA)	−0.05	0.01	−3.68	<**.001**
	Audio (EA–CA) × Video	−0.02	0.01	−1.09	0.28
	Audio (NA–EA) × Video	0.09	0.01	6.76	<**.0001**
	Audio (CA–NA) × Video	0.08	0.01	5.25	<**.0001**
	Location × Audio (EA–CA) × Video	0.11	0.03	3.79	<**.001**
	Location × Audio (NA–EA) × Video	−0.15	0.03	−5.57	<**.0001**
	Location × Audio (CA–NA) × Video	0.04	0.03	1.47	0.14
Total times	(Intercept)	5.75	0.03	201.26	<**.0001**
	Location (middle–end)	−0.02	0.04	−0.60	0.56
	Video (present–absent)	−0.24	0.02	−11.50	<**.0001**
	Audio (EA– CA)	0.05	0.02	2.74	**.01**
	Audio (NA–EA)	0.04	0.01	3.48	**.001**
	Audio (CA–NA)	−0.09	0.02	−4.88	<**.0001**
	Location × Video	0.10	0.01	8.20	<**.0001**
	Location × Audio (EA–CA)	−0.09	0.01	−6.14	<**.0001**
	Location × Audio (NA–EA)	0.05	0.01	3.53	**.0004**
	Location × Audio (CA–NA)	−0.04	0.01	−2.78	**.01**
	Audio (EA–CA) × Video	0.00	0.01	−0.07	0.95
	Audio (NA–EA) × Video	0.11	0.01	7.78	<**.0001**
	Audio (CA–NA) × Video	0.11	0.01	7.20	<**.0001**
	Location × Audio (EA–CA) × Video	0.11	0.03	3.86	**.0001**
	Location × Audio (NA–EA) × Video	−0.10	0.03	−3.78	**.0002**
	Location × Audio (CA–NA) × Video	−0.01	0.03	−0.28	0.78

Note. Bold font indicates $p < 0.025$. CA, Chinese audio; EA, English audio; NA, no audio.

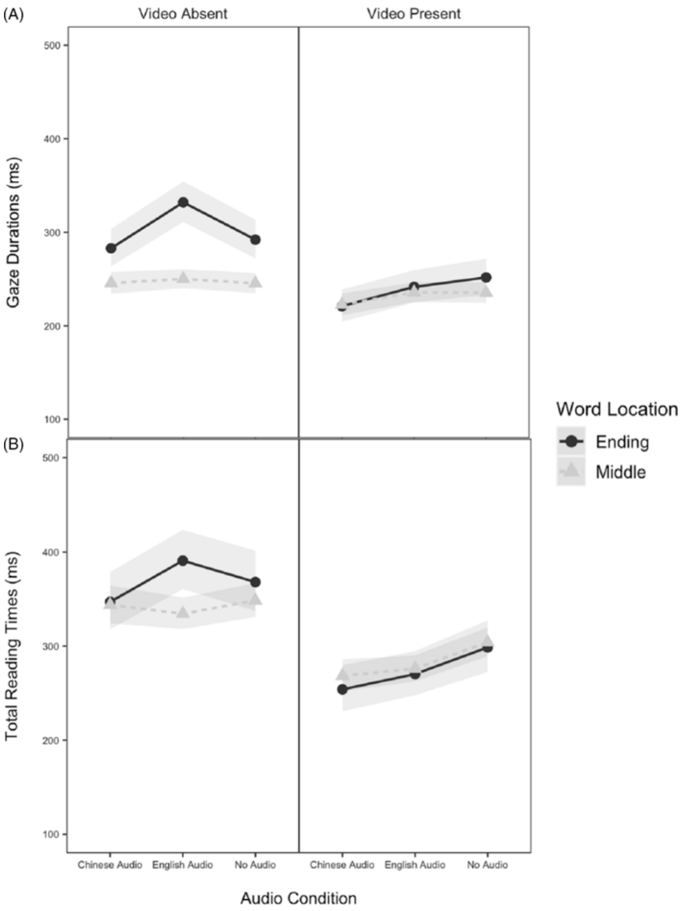


Figure 7. The LMMs-adjusted wrap-up effects as a function of video presence/absence and audio condition. Ribbons represent the lower (5%) and upper limits (95%) of confidence intervals for the estimated marginal means.

eye-movement measures provided supportive evidence for Hypothesis 1—that is, as the reading propensity decreased from no audio to English audio or Chinese audio, participants tended to rely less on subtitles, yielding fewer, shorter fixations, longer saccades on the subtitles (similar to the "skimming" pattern as observed in Liao et al.'s, 2021 study), as well as higher skipping rate of the subtitles.

While intelligible background speech has been generally found to disrupt the reading of static text by causing more fixations and longer reading times compared to reading in silence (see, e.g., Cauchard et al., 2012; Vasilev et al., 2019; Yan et al., 2018), such auditory disruption effect was not observed in our study. Instead, our results are in line with previous findings from subtitling research that auditory input could facilitate the reading of subtitles by, for example, reducing the reading times (see, e.g., d'Ydewalle et al., 1991; Szarkowska & Gerber-Morón, 2018) and skipping more subtitles (Ross & Kowler, 2013). However, it should be noted that the nature of the background speech used in our study is different from that of those used in reading research that caused the auditory disruption effect. Previous research on static text used speech that was unrelated to the text being read (Vasilev et al., 2019; Yan et al., 2018), or speech that was related to the written text but scrambled (e.g., Hyönä & Ekholm, 2016), whereas our study used semantically accurate and relevant speech. Processing the semantically irrelevant or meaningless (scrambled) speech may compete for the same resources as required for reading (e.g., resources for semantic or sentence processing; cf., Hyönä & Ekholm, 2016; Vasilev et al., 2019), thereby disrupting the reading process. On the contrary, because semantically relevant speech contains identical or similar meaning as in the subtitle, it therefore provides an additional source to establish a situational model that is the same or similar to the one developed by the reading of subtitles. This may explain why participants were still able to maintain some level of comprehension in the presence of Chinese or English audio even when subtitles were not processed as thoroughly as in the no-audio condition where reading subtitles was essential for the overall video comprehension.

It is worth noting that, although the "skimming"-like eye-movement patterns observed in the Chinese audio condition in the current study are similar to those observed in Liao et al.'s (2021) study when participants read the most rapidly displayed subtitles, the motivations for employing such strategy might be different. In Liao et al.'s (2021) study, skimming was probably motivated by increased task demands due to the limited availability of subtitles, whereas in the present study, skimming was probably adopted because thorough text processing becomes less compelling, or even unnecessary, with Chinese audio (i.e., the participants' native language) because this auditory information allows easy access to the same linguistic information (contained in the English subtitles) required to understand the video. In other words, while superficial reading in the former study seems more likely to reflect global task constraints beyond the reader's control, in the current study, it likely reflects a voluntary choice based on the reader's monitoring of the inherent trade-offs among different sources of information (e.g., audio vs. subtitle vs. video content) and strategic decisions about the relative importance of each.

Local (word-based) eye-movement analyses provided more clues about the strategies employed by participants to both monitor their needs for the subtitles and for adapting their reading behavior (eye movements) to these needs so as to effectively maximize their comprehension. Although participants spent less time on individual words when reading subtitles with Chinese audio than without audio, lexical processing of subtitles was not attenuated with Chinese audio, indicating that participants showed an equal level of sensitivity to lexical variables such as word frequency and word length. One possible reason is that, because the video content of our stimuli was simple and the subtitles were presented at a relatively slow speed, participants might have sufficient time to watch the video content and process the individual words in the non-native subtitles for language learning even though the subtitles were not essential for comprehension. However, the wrap-up effect was attenuated in Chinese audio compared to no audio (in the absence of concurrent video presentation), which likely reflects the fact that readers engaged in less (or more superficial) post-lexical processing of the subtitles whereby they need to integrate word meanings into sentences for the construction of a situational model that is essential for comprehension. Taken together, our Hypothesis 2 that lexical and

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

post-lexical processing of subtitles would be attenuated with Chinese audio than without audio was partially supported.

Similar to the Chinese-audio condition, participants with English audio also spent less time on individual words in the subtitle, but lexical processing was not attenuated. However, there was a more pronounced wrap-up effect with English audio compared to both the Chinese- and no-audio conditions, indicative of deeper post-lexical integration of word meanings within a clause/sentence representation in the former. This suggests that participants use English audio to support reading (or vice versa) to the extent that the audio permits the reader to engage more with the text and to process it more deeply—the type of processing that would otherwise be challenging when done alone (i.e., without audio) or unnecessary (i.e., with Chinese audio). These results partially supported Hypothesis 3, which predicted deeper lexical and post-lexical processing of the subtitles with English audio than without audio.

Finally, the consistent interaction between video condition and audio condition in the eye-movement measures demonstrated that the impact of audio was modulated by visual processing demands. In line with Hypothesis 4, while the overall pattern was one in which the presence of audio allowed viewers to rely less on the subtitles, the necessity or propensity to read the subtitles was further reduced with concurrent video content. Our separate manipulations of the video and audio conditions therefore show that, when situated in multimodal reading contexts containing multiple sources of information, readers can accurately gauge their needs for the subtitle and adjust their eye-movement routines to accommodate the task demands and maintain some desired level of comprehension.

It is also worth noting that, consistent with the findings reported by Liao et al. (2021), we observed an enhancing effect of video content—participants attained higher comprehension accuracy with concurrent video content than without, supporting *the multimedia principle* or notion that learning with text and picture is better than learning with text alone (Mayer, 2014). Moreover, the presence of video produced similar eye-movement patterns on subtitle reading as reported by Liao et al. (2021). Overall, with concurrent video content, participants made shorter, fewer fixations, and longer saccades on the subtitles. More subtitles were also skipped with video presentation than without. There could be two possible explanations for the fact that participants spent less time reading subtitles when video was present than absent. Participants might be attracted to the background video either because of the presence of dynamic visuals in the video (cf., exogenous influences on attentional selection during film viewing, Loschky et al., 2020) or because they strategically used the video content to support the reading and comprehension of subtitles (cf., the bottom-up vs. top-down control of attention, Awh et al., 2012).

Based on the finding that semantically relevant audio reduced the time spent on individual words of the subtitle but did not impair comprehension, we extended the multimodal integrated-language framework by adding more detailed assumptions about how auditory information (verbal and non-verbal) might influence overall comprehension and the reading of subtitles. New additions to the framework are presented inside the dashed-line box as shown in Figure 1. Like word or object identification, verbal input (e.g., spoken dialogue) in the audio can only be processed on a word-by-word basis because of its inherently serial nature. However, non-verbal input (e.g., background noise) can be processed in a parallel manner. Verbal input mainly

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

contributes to word processing by, for example, facilitating the identification of words and/or integration of words into larger linguistic units (e.g., clauses or sentences). Non-verbal input, on the other hand, is largely beneficial to the processing of non-textual visuals, such as the identification of an object or a scene where the event described in the text takes place. The facilitation of object/scene identification will in turn be conducive to the processing and comprehension of written text by allowing readers to use multiple sources to establish a more elaborate situation model.

To make these ideas more concrete, consider a specific hypothetical example of someone watching a video about someone playing with a cat with the subtitle "Sixin is playing with her cat." In this example (see Figure 1), the viewer receives visual inputs corresponding to the subtitles and other visual film elements (e.g., images of a girl and a cat), as well as the audio (e.g., the same sentence being spoken and other non-verbal sounds). Hearing the spoken word "cat" while simultaneously tracking the previously identified image of a cat on the screen would be expected to facilitate the identification of the referent's corresponding written form (i.e., the printed word "cat") in the subtitle, allowing its meaning to be retrieved from memory even under conditions where the printed form of the word may have been only superficial processed (e.g., fixated only briefly or identified from a distant viewing location). At the same time, other non-verbal auditory input (e.g., the "meow" sound of the cat) could also foster the identification of the cat in the scene, thereby contributing to the construction of a more elaborate situation model—one that is based on propositional representations generated by the processing of the subtitles, as well as propositional representation from other visual elements in the film.

In conclusion, by investigating the consequences of reading subtitles containing (partially) redundant auditory input and how this might modulate the high-level representations that people form from their video viewing experience, the current study provides important new information about multimodal reading. For example, our results provide clear evidence that eye movements in multimodal reading situations, such as reading subtitles in video, are not merely controlled by information from the visual modality. Instead, readers use inputs from both visual and auditory modalities in real time to make decisions about when and where (and even *if*) to move their eyes to read the subtitles. This complex decision making in turn indicates that the perceptual, cognitive, and oculomotor systems that are engaged during normal reading are both flexible and highly responsive to task demands. Moreover, eye-movement control during multimodal reading is much more complicated and nuanced than during "normal" reading (i.e., of statically displayed text), a complete understanding of which requires consideration of metacognitive strategies employed in evaluating the reader's need for subtitles to maintain effective comprehension (Andrews & Veldre, 2020). We admit that there are other factors that might modulate the effects of visual and auditory inputs (e.g., different strategies used in translating the audio into the subtitle might render different degrees of congruency in the semantic content between the two sources, thus affecting eye movements during subtitle reading, cf., Ghia, 2012; Ragni, 2020), but the study reported in this article brings us closer to understanding the mental processes underlying multimodal reading and the role of metacognition in these complicated visual-cognitive tasks.

Finally, despite its novel contributions, there are at least two limitations of the present study that need to be addressed in future research. First, comprehension

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

performance was tested using relatively simple multiple-choice questions, and therefore provides only limited insight into the influence of auditory information on high-level comprehension of the text. Second, the auditory processing was not measured in the present study. Future research might address this second limitation by, for example, using secondary auditory tasks to determine the extent to which attention is allocated to the processing of auditory input. Despite these limitations, however, our study clearly documents how the redundancy of auditory input both modulates the propensity to read subtitles and the eye-movement routines deployed to do so. By proceeding in this incremental fashion, we hope to provide a better understanding of one of the most complex activities that humans can engage in—the reading of subtitles in multimodal video contexts.

Financial support. The first author was supported by a PhD scholarship from Macquarie University.

Conflict of interest. The authors declared no conflict of interest concerning the authorship or the publication of this article.

Notes

[1] Software available at www.aegisub.org.

[2] In a small number of instances (13%) where one sentence spanned across two subtitles, the subtitle might not correspond perfectly to the audio due to subtitle break. This is inevitable because we used authentic materials in the market and it is challenging to achieve complete equivalence in information flow for English versus Chinese audios due to differences between these two languages (cf., Baker, 1992; Wu, 2010).

[3] As a sanity check for the possibility that longer reading times on subtitle-ending words might be driven by the visual boundaries in lines (i.e., people tend to slow down reading when approaching the end of a line) rather than by syntactic processing (Kuperman, Dambacher, Nuthmann, & Kliegl, 2010), we conducted a separate set of analyses using subtitles that were only sentences or clauses (32% data loss). The result patterns were the same only with numerical changes.

[4] Measures summarized in this table are based on the subtitle region unless otherwise indicated.

[5] The proportion of time spent on the subtitle/video as function of their presentation times.

References

Andrews, S. & Veldre, A. (2020). Wrapping up sentence comprehension: The role of task demands and individual differences, *Scientific Studies of Reading*, DOI: 10.1080/10888438.2020.1817028.

Atkinson, R. C. & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation*, **2**, 89–195.

Atkinson, R. C. & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American*, **225**, 82–90.

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: a failed theoretical dichotomy. *Trends in Cognitive Sciences*, **16**, 437–443.

Baker, M. (1992). *In other words: A course book on translation*. London and New York: Routledge.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, **68**, 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. Retrieved from https://arxiv.org/pdf/1506.04967.pdf.

Bisson, M. J., van Heuven, W. J. B., Conklin, K., & Tunney, R. J. (2014). Processing of native and foreign language subtitles in films: An eye tracking study. *Applied Psycholinguistics*, **35**, 399–418.

Cauchard, F., Cane, J. E., & Weger, U. W. (2012). Influence of background speech and music in interrupted reading: An eye-tracking study. *Applied Cognitive Psychology*, **26**, 381–390.

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

Cohn, N. (2016). A multimodal parallel architecture: A cognitive framework for multimodal interactions. *Cognition*, **146**, 304–323.

d'Ydewalle, G. & De Bruycker, W. (2007). Eye movements of children and adults while reading television subtitles. *European Psychologist*, **12**(3), 196–205.

d'Ydewalle, G., Praet, C., Verfaillie, K., & Van Rensbergen, J. (1991). Watching subtitled television: Automatic reading behavior. *Communication Research*, **18**(5), 650–666.

d'Ydewalle, G., Van Rensbergen, J., & Pollet, J. (1987). Reading a message when the same message is available auditorily in another language: The case of subtitling. In J. K. O'Regan & A. Lévy-Schoen (Eds.), *Eye movements: From physiology to cognition* (pp. 313–321). Amsterdam: North-Holland.

Fothergill, A. (Director). (2006). *Planet Earth*. Burbank, CA: BBC Video.

Ghia, E. (2012). The impact of translation strategies on subtitle reading. In E. Perego (Ed.), *Eye tracking in audiovisual translation* (pp. 157–182). Roma: Aracne.

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, **115**, 210–228.

Hirotani, M., Frazier, L., & Rayner, K. (2006). Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. *Journal of Memory and Language*, **54**, 425–443.

Hyönä, J., & Ekholm, M. (2016). Background speech effects on sentence processing during reading: An eye movement study. *PLoS ONE*, **11**, 1–25.

Inhoff, A. W. & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, **40**, 431–439.

Kruger, J.-L. (2016). Psycholinguistics and audiovisual translation. *Target*, **28**, 276–287.

Kruger, J.-L. & Doherty, S. (2016). Measuring cognitive load in the presence of educational video: Towards a multimodal methodology. *Australasian Journal of Educational Technology*, **32**, 19–31.

Kuperman, V., Dambacher, M., Nuthmann, A., & Kliegl, R. (2010). The effect of word position on eye-movements in sentence and paragraph reading. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, **63**, 1838–1857.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, **82**, 1–26.

Lång, J. (2016). Subtitles vs. narration: The acquisition of information from visual-verbal and audio-verbal channels when watching a television documentary. In S. Hansen-Schirra and S. Grucza (Eds.), *Eye tracking and Applied Linguistics* (pp. 59–82). Berlin: Language Science Press.

Liao, S., Kruger, J.-L., Doherty, S. (2020). The impact of monolingual and bilingual subtitles on visual attention, cognitive load, and comprehension. *The Journal of Specialised Translation*, **33**, 70–89.

Liao, S., Yu, L., Reichle, E. D., & Kruger, J.-L. (2021). Using eye movements to study the reading of subtitles in video. *Scientific Studies of Reading*, **25**, 417–435.

Loschky, L. C., Larson, A. M., Smith, T. J., & Magliano, J. P. (2020). The scene perception & event comprehension theory (SPECT) applied to visual narratives. *Topics in Cognitive Science*, **12**(1), 311–351.

Lüdecke, D. (2020). sjPlot: data visualization for statistics in social science. https://cran.r-project.org/web/packages/sjPlot/.

Marsh, J. E., Hughes, R. W., & Jones, D. M. (2008). Auditory distraction in semantic memory: A process-based approach. *Journal of Memory and Language*, **58**, 682–700.

Marsh, J. E., Hughes, R. W., & Jones, D. M. (2009). Interference by process, not content, determines semantic auditory distraction. *Cognition*, **110**, 23–38.

Martin, R. C., Wogalter, M. S., & Forlano, J. G. (1988). Reading comprehension in the presence of unattended speech and music. *Journal of Memory and Language*, **27**, 382–398.

Mayer, R. E. (2005). Cognitive Theory of Multimedia Learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimodal learning* (pp. 31–48). New York: Cambridge University Press.

Mayer, R. E. (Ed.). (2014). *The Cambridge handbook of multimedia learning*. New York: Cambridge University Press.

Paivio, A. (1986). *Mental representations: A dual coding approach*. New York: Oxford University Press.

Paivio, A. (2007). *Mind and its evolution: A dual coding theoretical approach*. Mahwah, NJ: Erlbaum.

Pollatsek, A., Juhasz, B. J., Reichle, E. D., Machacek, D., & Rayner, K. (2008). Immediate and delayed effects of word frequency and word length on eye movements in reading: A reversed delayed effect of word length. *Journal of Experimental Psychology: Human Perception and Performance*, **34**, 726–750.

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

Pylyshyn, Z. (2004). Some puzzling findings in multiple object tracking: I. Tracking without keeping track of object identities. *Visual Cognition*, **11**, 801–822.

Pylyshyn, Z. & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, **3**, 179–197.

Ragni, V. (2020). More than meets the eye: An eye-tracking study of the effects of translation on the processing and memorisation of reversed subtitles. *Journal of Specialised Translation*, **33**, 99–128.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, **124**, 372–422.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, **62**, 1457–1506.

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the E-Z Reader model. *Journal of Experimental Psychology: Human Perception and Performance*, **30**, 720–732.

Rayner, K., Kambe, G., & Duffy, S. A. (2000). The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology*, **53A**, 1061–1080.

Rayner, K. & McConkie, G. W. (1976). What guides a reader's eye movements? *Vision Research*, **16**, 829–837.

Reichle, E. D. (2021). *Computational models of reading: A handbook*. Oxford, UK: Oxford University Press.

Reichle, E. D., Liversedge, S. P., Pollatsek, A., & Rayner, K. (2009). Encoding multiple words simultaneously in reading is implausible. *Trends in Cognitive Sciences*, **13**, 115–119.

Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E-Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research*, **7**, 4–22.

Reichle, E. D., Pollatsek, A., & Rayner, K. (2012). Using E-Z Reader to simulate eye movements in nonreading tasks: A unified framework for understanding the eye-mind link. *Psychological Review*, **119**, 155–185.

Reichle, E. D., Yu, L., Liao, S., & Kruger, J.-L. (2021). Using simulations to understand the reading of rapidly displayed subtitles. *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*.

Ross, N. M., & Kowler, E. (2013). Eye movements while viewing narrated, captioned, and silent videos. *Journal of Vision*, **13**, 1–19.

Schilling, H. E. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition*, **26**, 1270–1281.

Schnotz, W. (2005). Integrated model of text and picture comprehension. In R. E. Mayer (Ed.), *Cambridge handbook of multimodal learning* (pp. 49–70). New York: Cambridge University Press.

Stowe, L. A., Kann, E., Sabourin, L., & Taylor, R. C. (2018). The sentence wrap-up dogma. *Cognition*, **176**, 232–247.

Szarkowska, A. & Gerber-Morón, O. (2018). Viewers can keep up with fast subtitles: Evidence from eye movements. *PLoS ONE*, **13**, 1–30.

Tiffin-Richards, S. P. & Schroeder, S. (2018). The development of wrap-up processes in text reading: a study of children's eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **44**, 1051–1063.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, **67**, 1176–1190. DOI: 10.1080/17470218.2013.850521.

Vasilev, M. R., Kirkby, J. A., & Angele, B. (2018). Auditory distraction during reading: A Bayesian meta-analysis of a continuing controversy. *Perspectives on Psychological Science*, **13**, 567–597.

Vasilev, M. R., Liversedge, S. P., Rowan, D., Kirkby, J. A., & Angele, B. (2019). Reading is disrupted by intelligible background speech: Evidence from eye-tracking. *Journal of Experimental Psychology: Human Perception and Performance*, **45**, 1484–1512.

von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, **94**, 119–133.

Warren, T., White, S.J., & Reichle, E. D. (2009). Investigating the causes of wrap-up effects: Evidence from eye movements and E-Z Reader. *Cognition*, **111**, 132–137.

Wu, G. (2010). Translation difference – A hybrid model for translation training. *The International Journal for Translation & Interpreting Research*, **2**, 24–37.

https://doi.org/10.1017/S0142716421000527 Published online by Cambridge University Press

Yan, G., Meng, Z., Liu, N., He, L., & Paterson, K. B. (2018). Effects of irrelevant background speech on eye movements during reading. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, **71**, 1270–1275.

Zhang, H., Miller, K., Cleveland, R., & Cortina, K. (2018). How listening to music affects reading: Evidence from eye tracking. *Journal of Experimental Psychology: Learning Memory and Cognition*, **44**, 1778–1791.

Appendix 1. Summary of final models in data analyses.

Measures	Final models	Total number of observations	Marginal R^2 (conditional R^2)
Comprehension	Video*Audio + (1 | Subject) + (1 | Video Item)	1632	0.01 (0.06)
Average fixation durations	Video*Audio + (Video*Audio | Subject) + (Video | Subtitle)	13172	0.04 (0.36)
Total number of fixations	Video*Audio + (Video*Audio | Subject) + (Video | Subtitle)	13879	0.11 (0.64)
Progressive saccade length	Video*Audio + (Video + Audio | Subject) + (1 | Subtitle)	52146	0.01 (0.08)
Percentage of skipped subtitles	Video*Audio + (Video + Audio | Subject) + (Video | Subtitle)	13879	0.20 (0.69)
Word-frequency and length effects (gaze durations)	Frequency*Length*Video*Audio + (Video + Audio + Frequency + Length | Subject) + (Video | Word)	28620	0.12 (0.22)
Word-frequency and length effects (total times)	Frequency*Length*Video*Audio + (Video + Audio + Frequency + Length | Subject) + (Video + Audio | Word)	28620	0.26 (0.38)
Wrap-up effect (gaze durations)	Word Location* Video*Audio + (Video + Audio + Word Location | Subject) + (Video | Word)	39295	0.03 (0.23)
Wrap-up effect (total times)	Word Location* Video*Audio + (Video + Audio + Word Location | Subject) + (Video | Word)	39295	0.04 (0.37)

Note. Number of subtitle items for all global analyses: 455; number of word items for word-frequency and word-length effects: 1919; number of word items for the wrap-up effect: 2428. Marginal R^2 evaluates the variance explained by the fixed effects, while conditional R^2 evaluates the variance explained by both fixed and random effects. Marginal R^2 and conditional R^2 were produced using *sjPlot* package (Lüdecke, 2020).

Cite this article: Liao, S., Yu, L., Kruger, J-L., and Reichle, ED. (2022). The impact of audio on the reading of intralingual versus interlingual subtitles: Evidence from eye movements. *Applied Psycholinguistics* **43**, 237–269. https://doi.org/10.1017/S0142716421000527