



A synthesis of research on L2 vocabulary learning through audiovisual input and on-screen text

INJUNG WI (Corresponding author)  

FRANK BOERS  

Faculty of Education, University of Western Ontario

Abstract

Given the cognitive and pedagogical affordances of audiovisual materials and ongoing technological advancements, there is an increasing need to explore how they can be optimized for effective second language (L2) learning (Montero Perez, 2022). The multimodal nature of audiovisual input, particularly when enhanced with on-screen text, offers significant potential for developing multifaceted vocabulary knowledge. This narrative review synthesizes key theoretical frameworks, including Mayer's Cognitive Theory of Multimedia Learning (Mayer, 2021), Krashen's Input Hypothesis (Krashen, 1982), Vygotsky's Zone of Proximal Development (Vygotsky, 1978), and Bjork's Desirable Difficulties (Bjork, 1994), alongside recent empirical findings. We examine the pedagogical potential of various types of on-screen text (e.g., L1 captions, L2 captions, and bilingual subtitles), identify gaps in the existing literature, discuss methodological challenges and highlight promising directions for future research. By doing so, this review seeks to deepen our understanding of how audiovisual input enhanced with on-screen text can support the development of L2 vocabulary knowledge in an increasingly digital learning landscape.

Keywords: L2 vocabulary learning; audiovisual input; multimodal input; on-screen text; captions; subtitles; dual subtitles; typographic enhancement; glossing

Introduction

Audiovisual input has emerged as a valuable source of authentic spoken language for second language (L2) learners. With the advancement of Information and Communications Technology (ICT), video content is now widely accessible through platforms such as YouTube, Netflix, Disney+, and TED Talks, making it a popular tool for language learning. Unlike auditory input alone (e.g., radio or audiobooks), audiovisual materials offer dual channels, audio and visual,

providing richer, multimodal learning experiences. This dual-modality benefit is supported by Mayer's Cognitive Theory of Multimedia Learning, which posits that verbal and non-verbal information are processed through separate but interconnected cognitive systems (Mayer, 2021). When learners receive input through both auditory and visual channels, this may promote understanding and consequently retention of the information they have successfully processed.

An additional layer of multimodal support comes from the integration of on-screen text, such as subtitles or captions.¹ On-screen text can appear in the learner's first language (L1) as subtitles, in the target language (L2) as captions, or as bilingual (or dual) subtitles that display both. Numerous studies have shown that this textual support enhances the comprehensibility of audiovisual input, making it more beneficial for L2 learners (Montero Perez et al., 2013; Winke et al., 2010). The importance of comprehensible input for learning has long been recognized, most explicitly by Krashen's Input Hypothesis (Krashen, 1982), but also less directly by notions such as Vygotsky's Zone of Proximal Development (Swain et al., 2015; Vygotsky, 1978, 1998) and Desirable Difficulties (Bjork, 1994, 2018). These frameworks suggest that learning occurs most effectively when the learning materials pose challenges but are nonetheless accessible, a balance that on-screen text can help achieve by scaffolding learners' understanding of audiovisual materials.

As a form of multimodal scaffolding, on-screen text plays a significant role in enhancing various aspects of L2 acquisition. Research in Second Language Acquisition (SLA) in general and Computer-Assisted Language Learning (CALL) more specifically has documented its facilitative effects on listening comprehension (e.g., Kang, 2019; Montero Perez et al., 2014a; Yang & Chang, 2014), vocabulary development (e.g., Danan, 1992, 2004; Frumuselu et al., 2015; Hsieh, 2020; Kurokawa et al., 2024; Casulleras & Miralpeix, 2024; Peters & Webb, 2018; Teng & Cui, 2025; Wi & Boers, 2024), grammar acquisition (Lee & Révész, 2018, 2020; Muñoz et al., 2023; Pattermore et al., 2024), pronunciation (Mitterer & McQueen, 2009; Mora & Fouz-González, 2024; Wong et al., 2020), and even writing instruction (Alobaid, 2021). Among these, vocabulary learning has been the most widely studied area, with a growing body of evidence suggesting that viewing video with L1 or L2 captions leads to significantly better vocabulary gains compared to viewing without textual support.

The positive impact of on-screen text on vocabulary acquisition is especially promising given the multidimensional nature of vocabulary knowledge. Apart from semantic knowledge, knowing a word includes how it is written and pronounced, what type of discourse it belongs to, and what its usual grammatical and phraseological behavior is. Nation (2022) outlines the multiple dimensions of vocabulary knowledge, broadly categorized into form (written and spoken), meaning, and use. It is reasonable to expect that audiovisual input enhanced with on-screen text can better support not only the acquisition of form-meaning connections, but also phonological-orthographic linking and phonological-morphological awareness, thanks to the complementary channels of information (i.e., text and audio). This distinguishes captioned audiovisual input from monomodal input, such as reading or listening. Arguably, reading *while* listening to an audio-recording of the text provides the same learning opportunities regarding the written and spoken forms of words, but we shall see that the video component of audiovisual input can offer additional benefits.

Given the cognitive and pedagogical affordances of audiovisual materials and continually developing technology, there is a growing need to further explore how they can best be utilized

¹ In this review, we use the terms (*L1*) subtitles and *L1* captions interchangeably, reflecting the varying terminology employed by different authors in the existing literature.

for effective L2 vocabulary acquisition. In this narrative review, we synthesize key theoretical frameworks and empirical findings related to video-based L2 learning supported by on-screen text, with a particular focus on vocabulary acquisition. We will explore the pedagogical potential of L1 captions and L2 captions (and their combination), identify gaps in current research, and highlight promising directions for future studies. By doing so, this review aims to deepen our understanding of how audiovisual input can be optimized to support the development of L2 vocabulary knowledge in an increasingly digital learning landscape.

Theoretical Framework

This section reviews two key theoretical perspectives that explain why video-based multimodal input, particularly when enhanced with on-screen text, can support second language (L2) vocabulary acquisition. The first is Mayer's Cognitive Theory of Multimedia Learning, which offers an account of how learners process verbal and visual information. The second perspective combines Krashen's Input Hypothesis, Vygotsky's Zone of Proximal Development, and Bjork's Desirable Difficulties and emphasizes the critical role of balancing task difficulty and task feasibility in facilitating L2 learning. Together, these perspectives justify integrating on-screen text with audiovisual input to support vocabulary learning in authentic and challenging contexts.

Cognitive Theory of Multimedia Learning

The Cognitive Theory of Multimedia Learning (CTML), developed by Mayer (1997), proposes that learners understand information more effectively when it is presented through both words and pictures than through words alone. The term "words" encompasses both spoken and written forms (e.g., narration or on-screen text), while "pictures" include static visuals (e.g., illustrations, charts) and dynamic media (e.g., videos or animations) (Mayer, 1997, 2014, 2021; Mayer & Moreno, 2003). As demonstrated by Mayer and associates (Mayer & Anderson, 1992; Mayer & Gallini, 1990), visuals can be very helpful, for example, to promote comprehension of scientific texts and of technical manuals, where verbal explanations may not suffice to help readers 'picture' the described concepts or processes.

The theory is based on three fundamental assumptions: (1) the dual channels assumption, (2) the limited capacity assumption, and (3) the active processing assumption (Mayer, 2017, 2021). The dual channels assumption posits that the human mind has two separate channels for processing verbal and visual information. This concept is reminiscent of Paivio's Dual Coding Theory (Paivio, 1990, 2007, 2010), which distinguishes between concepts that are represented verbally in the mind and concepts that are, in addition, represented in a non-verbal way (as mental images and other ways directly linked to the senses). The verbal and non-verbal coding systems can operate separately or in tandem. When knowledge is represented in the mind in both verbal and non-verbal modes, it is easier to recall according to the theory. In essence, the theory explains why concrete concepts are generally easier to retrieve from the mental lexicon than abstract ones—unless a way is found to make the latter imageable as well (e.g., Farley et al., 2012; Shen, 2010).

The second assumption of CTML, limited capacity, asserts that each channel can process only a limited amount of information at one time (Mayer, 2021). This view aligns with Sweller's Cognitive Load Theory (Sweller et al., 2011), which holds that working memory is restricted in capacity and duration. This second assumption is important because it indicates that multimodal input is not always beneficial but needs to be designed with consideration for learners' processing capacity. Overloading the verbal or visual channel (or both) can hinder learning. A study by Ariew and Erçetin (2004), for example, examined whether providing video clips to supplement the content of a reading text supported L2 learners' comprehension of this

text. They found a negative effect, indicating that the processing demands of the video clips may have come at the expense of the attention that the learners would otherwise have given to the actual text. However, when the information of two channels of input is congruent, then learners do appear to benefit from multimodal input. For example, Lee and Mayer (2015) found that adding video to an audio-only lesson significantly improved L2 learners' comprehension of the lesson content. Additionally, the learners in the audio + video condition reported less difficulty during learning than their peers who used the audio-only material.

The third assumption of CTML, active processing, posits that learners must actively engage with the material to construct meaningful mental representations. Multimodal input invites learners to select and combine relevant verbal and visual information, organize it into coherent mental structures, and integrate it with prior knowledge (Mayer, 2021). In this context, it is worth noting that Moreno's Cognitive-Affective Theory of Learning with Media (CATLM; Moreno & Mayer, 2007), which builds on CTML, cautions that active or generative processing may not occur if learners are not sufficiently motivated or interested in the material (also see Vanderplank's cognitive-affective model of language learning through captioned viewing; Vanderplank, 2016; Teng & Vanderplank, 2014). These theories emphasize the important role of motivational factors for successful multimedia learning.

CTML offers a robust foundation for understanding how audiovisual materials with on-screen text can support L2 vocabulary acquisition. These materials combine verbal (audio and text) and visual (images, gestures, and facial expressions) modalities, enabling learners to process language through multiple channels. However, it could be argued that presenting learners with identical content through two or more channels is unnecessary if one channel suffices for them to comprehend this content. For instance, if pictures are sufficient to follow the instructions in a manual, then adding verbal descriptions may be redundant. This redundancy may also apply to the addition of on-screen text in the case of audiovisual materials (movies, TV shows, etc.). Unless they have hearing problems, English native speakers may experience English captions as redundant, as they have no problem comprehending the spoken language. In this context, it is important to situate the topic of this article in the realm of L2 acquisition, where learners can find understanding spoken discourse very challenging. For example, in a study by Lee and Mayer (2018), Korean university students who watched an English-language video with English captions outperformed those who watched the video without the captions on a comprehension test. This finding suggests that, for L2 learners, on-screen text can serve as valuable support to help them manage the cognitive demands of processing fast-paced or unfamiliar spoken language.

Theories Emphasizing Comprehensible Input

Another foundational perspective supporting the use of on-screen text in L2 learning is the concept of comprehensible input. When learners cannot comprehend the input, opportunities for learning new vocabulary are diminished. As we shall see below, on-screen text can serve as a scaffolding tool, giving learners access to content that might otherwise be incomprehensible. Several theories emphasize the importance of this type of support. First, Krashen's Input Hypothesis (Krashen, 1982; Krashen & Terrell, 1983) argues that language acquisition occurs when learners are exposed to input slightly beyond their current competence, referred to as $i+1$. If input is too simple, it offers little challenge; if it is too difficult, it may be incomprehensible. Optimal learning occurs when the input is just challenging enough to promote growth. On-screen text helps maintain this balance by making difficult input more accessible.

Second, Vygotsky's Zone of Proximal Development (ZPD) offers a sociocultural perspective on scaffolding. ZPD refers to the gap between what learners can achieve independently and what

they can accomplish with support from a more knowledgeable other or a mediating tool (Swain et al., 2015). On-screen text functions as such a tool, enabling learners to engage with aural input that would otherwise be beyond their reach.

Third, the concept of Desirable Difficulties, from the domain of cognitive psychology, reinforces this idea. It suggests that learning tasks should be challenging enough to promote effortful processing but not so difficult that learners fail (Bjork, 1994, 2018; Schmidt & Bjork, 1992). If a learner cannot respond to the difficulty successfully, learning becomes undesirable and less likely to occur. On-screen text reduces the risk of failure by supporting comprehension and enabling more effective engagement with the input. At the same time, L2 learning through videos can be rendered desirably difficult by calibrating the level of challenge, for example, by using L2 captions instead of L1 subtitles or by gradually reducing the amount of textual support as learners' proficiency increases. These adjustments help maintain an optimal balance between support and challenge, promoting deeper processing and sustained language development. Although these theories originate from different disciplinary backgrounds, they share a central tenet: learning occurs most effectively when input is comprehensible yet sufficiently challenging.

In sum, both cognitive and pedagogical theories support the use of audiovisual input enhanced with on-screen text in L2 learning. Mayer's CTML explains how learners process multimodal input, while the theories emphasizing comprehensible input show why such support is often necessary for language acquisition. Together, they provide a comprehensive theoretical basis for the instructional use of captioned or subtitled videos in L2 vocabulary learning.

Literature Review

Vocabulary Learning through Audiovisual Input

Vocabulary knowledge is widely recognized as a fundamental component of L2 proficiency. Repeated exposure to words in meaningful contexts contributes to vocabulary development. Studies have shown that content-focused activities such as reading (e.g., Horst et al., 1998; Pellicer-Sánchez & Schmitt, 2010; Webb & Chang, 2015), listening (van Zeeland & Schmitt, 2013; Vidal, 2011), reading while listening (e.g., Brown et al., 2008; Teng, 2018; Valentini et al., 2018), and viewing (e.g., Dang et al. 2022; Montero Perez, 2020; Neuman & Koskinen, 1992; Rodgers & Webb, 2011) provide opportunities for L2 lexical development. Until recently, unimodal input, notably reading (written input) and listening (auditory input), was the most common type of material that L2 learners used both inside and outside the classroom. However, multimodal input (e.g., audiovisual input) has over time become increasingly prominent for both educational and recreational purposes (Mayer, 2017; OECD, 2009; Statistics Canada, 2019; Vanderplank, 2019; Zenith, 2019).

In recent years, a growing body of research has highlighted the advantages of audiovisual input for L2 vocabulary learning. First and foremost, many audiovisual materials are motivating and engaging (Baranowska, 2020; Baltova, 1994). Second, they tend to offer rich aural input that mirrors real-life language use (Danan, 2004; Webb, 2015) and may provide repeated exposure to the same lexical items (Rodgers & Webb, 2011; Webb & Rodgers, 2009). Third, the visual and contextual cues, such as images, facial expressions, and gestures, support learners in inferring the meanings of unfamiliar words (Danan, 2004; Peters, 2019; Rodgers, 2018).

Empirical studies have shown that L2 vocabulary can be acquired from audiovisual input even if no subtitles or captions are available (e.g., Montero Perez, 2020; Neuman & Koskinen, 1992; Peters & Webb, 2018; Rodgers & Webb, 2020). For instance, Montero Perez (2020) found that learners who watched a 25-minute French documentary containing pseudowords outperformed a control group on form and meaning recognition, though no gains were observed

in meaning recall. Similarly, Puimège and Peters (2019) reported form recall of single words and multiword expressions (MWEs) from a reality TV clip, although meaning recall remained limited, and results were cautiously interpreted due to the small sample size. Peters and Webb (2018) examined the impact of a full-length (1-hour) BBC documentary. While form recognition results were inconclusive, learners showed modest gains in meaning recall, averaging four out of 64 words learned. Finally, Rodgers and Webb (2020) explored extensive viewing, with 187 students watching 10 episodes of a TV series. Despite the considerable input, learners acquired only about six of the 60 target words. A more recent study, by Dang et al. (2022), reported similar results with meaning-recall rates of 8.14% (about 4 out of 50 words). Overall, these findings suggest that, while vocabulary learning can occur through content-focused engagement with audiovisual input, the gains tend to be very modest. The inclusion of on-screen text, which we turn to next, may improve the learning outcomes.

Use of On-screen Text

L2 Captions

A substantial body of research has examined the facilitative role of on-screen text, particularly L2 captions, in L2 vocabulary acquisition. Although both L2 captions and L1 subtitles support learning (Peters et al., 2016), L2 captions or intralingual subtitles, which present text in the same language as the audio, have received most attention from researchers (Hsieh, 2020; Kurokawa et al., 2024; Montero Perez, 2022; Teng & Cui, 2025). While originally developed to assist the deaf and hard of hearing, captions were introduced into language classrooms in the 1980s to attract students' attention, lower their affective filters, and enhance their motivation (Winke et al., 2010; Vanderplank, 1988). Research shows that captions help learners visualize phonological forms (i.e., match the spoken words to their written forms). This helps learners to eventually identify spoken forms without textual support (Bird & Williams, 2002; Charles & Trenkic, 2015). Additionally, captions aid speech segmentation, that is, they help learners break up continuous speech into meaningful lexical units, an essential skill for vocabulary learning (Charles & Trenkic, 2015; Vanderplank, 1988). Collectively, these findings highlight the usefulness of L2 captions to scaffold vocabulary development during video-based learning.

Montero Perez et al. (2013) conducted a meta-analysis to examine the overall efficacy of L2 captioned video for listening comprehension and vocabulary acquisition. They analyzed 15 studies on listening comprehension and 10 on vocabulary learning, and concluded that captions had a generally positive effect on both outcomes. Kurokawa et al. (2024) conducted a more recent meta-analysis that updated Montero Perez et al.'s (2013) findings by incorporating a broader range of studies (49 studies) on incidental vocabulary acquisition through captioned videos. Like the earlier analysis, they found that L2 captions had a facilitative effect on vocabulary learning. However, the effect size reported was moderate ($g = 0.56$), compared to the larger effect ($g = 0.87$) found in the 2013 study. Kurokawa et al. (2024) suggested that this difference may be due to their exclusive focus on incidental learning, whereas Montero Perez et al. (2013) included studies involving intentional learning conditions. Furthermore, only four studies overlapped between the two meta-analyses, which may also contribute to the differing results.

In addition to standard L2 captions (i.e., verbatim transcriptions of the audio), the benefits of modified versions of captions have been investigated (Montero Perez, 2022). Informed by studies of L2 vocabulary acquisition from reading texts, these modified caption types include typographically enhanced items (where the selected items are made more salient by means of underlining, highlighting, or bolding), keyword captions (which display only the most essential

words), and glossed captions (which provide L1 translations for selected words). Montero Perez et al. (2014b) examined the effects of these captioning methods on L2 French learners by testing four viewing conditions: no captions, full captions, keyword captions, and full captions with highlighted key items. The results showed that all captioned conditions outperformed the no-caption condition in a form-recognition test. In the meaning-recognition test, only the keyword and enhanced caption conditions yielded significant improvements. However, meaning recall scores were uniformly low, with no significant differences across conditions. The study also found no significant effects on content comprehension. These findings suggest that, while modified captions aid recognition of word forms and meanings, they may not be sufficient to support deeper semantic processing. Similarly, Montero Perez et al. (2015) found that keyword captions produced better outcomes than full captions in a form-recognition test. However, consistent with the findings of Montero Perez et al. (2014b), recall remained poor across all conditions.

Unlike the previous studies, Montero Perez et al. (2018) included glossed keyword captions to examine the relative effectiveness of different captioning modes: no captions, full captions, keyword captions, and glossed keyword captions. Glossed captions provided learners with context-bound L1 translations of target words, which could be accessed by tapping the spacebar, which paused the video and displayed the glossed item in a centered pop-up box. The study found that glossed keyword captions led to significantly higher learning gains at the meaning recall level than both full and keyword captions. These findings support previous research showing that while enhanced and keyword captions aid form recognition, they may not suffice for semantic learning. The results suggest that explicit L1 translations (or other means of clarification) may be necessary when the meanings of unfamiliar words cannot be inferred from contextual or visual cues.

Hsieh (2020) also examined the differential effects of caption types on vocabulary learning and comprehension. The following viewing conditions were examined: without captions, with captions but without the audiotrack of the video, with standard captions, with captions in which the target words were highlighted, and with captions where these highlighted target words were accompanied by L1 glosses. Participants were low-intermediate Chinese speaking learners of English who watched two TED-Ed videos twice. After the second viewing, a video-comprehension test and three sets of vocabulary tests (form recognition, meaning recognition and meaning recall) were administered. The findings showed that the group who watched with enhanced captions with L1 glosses performed better on the form recognition test than the no-captions group and the group that watched the video with full captions but no audio, and significantly outperformed the other four groups on the meaning recognition and recall tests. These findings clearly indicate that captions can more effectively promote semantic lexical knowledge if learners are given explicit access to word meanings. Nonetheless, there was no significant difference in comprehension among the caption types. It should be noted that the test items were presented only through written prompts, which may have disadvantaged the no-captions group. Majuddin et al. (2021) compared the effects of no captions, full captions, and typographically enhanced captions on learners' acquisition of multiword expressions (MWEs) from watching a sitcom episode. They also examined if watching the same video twice instead of only once had an effect on the learning of these expressions. Their form recall test results showed that both full and enhanced captions produced better outcomes than the no-caption condition, and repeated viewing further improved the learning of MWEs. However, the typographic enhancement did not result in significantly greater gains than full captions. This contrasts with earlier studies (e.g., Montero Perez et al., 2014b, 2015, 2018). A potential explanation lies in the length of the target MWEs, with some containing up to five words, which

may have reduced the salience effect of enhancement due to the transient nature of spoken input (Montero Perez, 2022). In addition, the authors recognized that, even if typographic enhancement effectively increases the visual salience of items, it does not directly support meaning construction (Majuddin et al., 2021, p. 994). Majuddin et al. (2024) report a replication of their 2021 study, with the same group of students as in the initial study, but this time the students were aware of the upcoming types of vocabulary tests. The viewing conditions (no captions, captions, and enhanced captions) and the number of viewings (once or twice) were the same as in the previous study. Results showed that both caption conditions outperformed the no-caption condition, with the group that watched the video twice with enhanced captions achieving the highest gains in both immediate and delayed posttests. While the findings generally align with those of the original study, a notable difference emerged: enhancement was a stronger predictor of learning outcomes than repeated viewing. This suggests that the effect of enhancement is amplified under intentional learning conditions. Worth mentioning as well is that the attention that the students gave to the typographically enhanced items may have come at the expense of their processing of the content of the videos, because their scores on content-related quizzes were poor in comparison to the viewing condition with standard captions. This trade-off effect has also been attested when typographic enhancement is applied to reading texts (e.g., Boers et al., 2017; Choi, 2017), but the effect is probably more pronounced in the case of audiovisual materials, where learners have very limited time to process each caption and do not revisit earlier parts of the video to verify their comprehension. It has been suggested (e.g., Han et al., 2008) that learners could first engage with an unmodified version of a text to ensure they process its content, and subsequently present the same text in which a certain language feature or certain language items are visually enhanced. The second time the learners read the same text, they will already be familiar with its content, and so they will be able to allocate more of their attentional resources to the wording used to express the content. This principle applies to repeated reading (or listening) in general, but the advantage of adding visual enhancement when a text is processed a second time is that it harnesses the freed-up attentional resources by guiding learners' attention to the specific patterns or items that the materials designer hopes they will take in.

Puimège et al. (2023) also investigated the effects of unenhanced and enhanced L2 captions (with underlined MWEs) on form recall, supplemented by eye-tracking data. Thirty Flemish students viewed a 30-minute documentary containing 22 target MWEs, including idioms (e.g., *beyond the pale*), collocations (e.g., *supernatural powers*), compounds (e.g., *guinea pig*), and phrasal verbs (e.g., *tap into*). Enhanced captions led to greater form recall than unenhanced ones, and eye-tracking confirmed that participants fixated significantly longer on underlined target items. Despite this attention-directing effect, the authors emphasized that the effectiveness of typographic enhancement may vary depending on item difficulty and cognitive processing demands. It would be interesting to replicate this study but have the learners watch the same video twice, to examine if content familiarity after a first viewing enables learners to engage better with the target MWEs. In addition, it would be interesting to replicate studies such as this by presenting learners with standard captions in the first viewing and add the visual enhancement only on the second viewing, as suggested previously.

Unlike the studies mentioned above that involved adult learners, Finger-Bou and Muñoz (2024) investigated incidental vocabulary learning by young learners through extensive viewing with either full captions or textually enhanced captions. Seventeen fifth-grade students watched 11 episodes of an animated TV series. The findings revealed that students in the enhanced captions group outperformed those in the unmodified captions group on both form recognition and form recall tests. However, it remains unclear whether textually enhanced captions also

facilitate the acquisition of semantic knowledge. Although typographic means can make items more prominent, they may not help learners infer the meanings of these items. It is worth mentioning that some studies (e.g., Vu & Peters, 2020) which used reading-only input did find a positive effect of typographic enhancement even at the level of semantic knowledge. However, when learners' curiosity about an enhanced word or phrase is piqued during self-paced reading, this may prompt the learner to more carefully examine the item's broader context to try and figure out its meaning. This is less feasible in the case of captioned video, unless the learner decides it is worth pausing the video, and winding back and forth to inspect elements in the broader context beyond the caption line that includes the enhanced item.

L1 Captions (Subtitles)

Like L2 captions, L1 captions, also known as interlingual subtitles, have been found to support both content comprehension and L2 vocabulary acquisition (Danan, 2004; Wi & Boers, 2024). These subtitles provide L1 translated text alongside L2 audio. Viewing with L1 captions creates a complex processing task that invites learners to associate visual imagery, L2 speech, and L1 text simultaneously (Danan, 1992, 2004; Peters et al., 2016). While L1 captions may divert attention from the L2 soundtrack and may in fact make the latter redundant for comprehension of the video content, research suggests that automatic reading of subtitles does not necessarily hinder auditory encoding (Danan, 2004; d'Ydewalle & Pavakanun, 1997; Koolstra & Beentjes, 1999). In the study by Koolstra and Beentjes (1999), 246 Dutch primary school students watched a 15-minute episode of an English wildlife documentary, either with Dutch subtitles or without any subtitles. This study specifically included a (spoken) word recognition test to examine whether the presence of subtitles would interfere with children's ability to hear and recognize the English words. The results showed that the students who had watched the subtitled version in fact recognized the spoken forms of the English words more accurately than those who had watched the documentary without subtitles. The authors suggested that the spoken forms of new words can be more effectively processed through the two channels of input (listening and reading) and that L1 subtitles do not hinder but rather facilitate the recognition of spoken L2 words. It needs to be said here that Dutch and English are closely related languages that share a lot of phonological properties, share an alphabetic script, and have many cognates. It is uncertain whether Koolstra and Beentjes' findings are generalizable to other L1-L2 combinations.

In a study by Krugal et al. (2014), university students whose first language was Sesotho watched a 14-minute English psychology lecture under three conditions: no captions, L1 captions (Sesotho), and L2 captions (English). Comprehension of the lecture content was assessed immediately and after a two-week delay. While no significant differences emerged in the immediate test, the condition with L1 captions brought about the best performance in the delayed comprehension test. Additionally, participants who viewed the L1-captioned version reported lower cognitive effort and higher engagement, whereas the L2 captions group experienced the highest cognitive load. These findings highlight the potential of L1 captions to reduce processing demands while supporting comprehension. Similarly, Lee et al. (2025) examined the relationship between caption type (L1 vs. L2) and cognitive load using both electroencephalography (EEG) and a self-report questionnaire. Korean university students were assigned to one of three conditions while watching six videos: a baseline condition (Arabic, unfamiliar language), L1 captions (Korean), or L2 captions (English). Both the EEG data and questionnaire responses indicated that L1 captions greatly reduced cognitive load, whereas L2 captions led to only a moderate reduction in cognitive load compared to the baseline condition. The L1 captions were also found to be the most effective at helping the participants comprehend

the videos. In short, L1 captions can support comprehension and alleviate cognitive demands, which, in turn, may enable learners to allocate attentional resources to the L2 soundtrack.

Despite these benefits, fewer experimental studies have explored the potential of L1 captions for vocabulary learning compared to the more extensive research on L2 captions, and yet it stands to reason that L1 captions can help learners not only understand the content of videos but also the meaning of new L2 words. For example, the aforementioned study by Koolstra and Beentjes (1999) found that children in the L1 subtitled condition significantly outperformed those without subtitles not only on a (spoken) word recognition test but also a meaning recognition test.

However, a larger body of research has focused on comparing the benefits of L1 and L2 captions. In Birulés-Muntané and Soto-Faraco (2016), Spanish-speaking learners of English watched an English TV drama under three conditions: L1 captions, L2 captions, or no on-screen text. Although no differences were found in vocabulary learning gains between the captioned conditions, L2 captions led to better scores on a listening test designed to assess auditory perception (participants completed a gap-fill task by writing the missing words after hearing the corresponding spoken passage twice), while L1 subtitles performed significantly better on the comprehension test about the plot of the TV drama. Pujadas and Muñoz (2019) examined the effects on vocabulary learning of L1 vs. L2 captions, pre-teaching, and learner proficiency among Catalan-Spanish bilingual adolescents. Learners who watched the TV series with L1 captions performed better on a meaning-recall test, but no difference emerged in a form-recall test. In contrast, Peters (2019) found an advantage for L2 captions over L1 subtitles and no on-screen text on both form-recognition and meaning-recall measures. Baranowska (2020) investigated the effects of L1 and L2 captions on cognitive load, comprehension, and vocabulary learning among Polish high school students. While the students in the L2 caption condition gained the most vocabulary, L1 captions required less cognitive effort and led to better comprehension of the video content. It is worth mentioning that the group of students assigned to the L1 caption condition happened to have lower vocabulary pre-test scores, with the difference between the two groups approaching statistical significance ($p = .07$). This may have put them at a disadvantage. For example, Muñoz et al. (2024) found that pre-test scores were a significant predictor of vocabulary acquisition from captioned video. A recent study by Casulleras and Miralpeix (2024) conducted a longitudinal study to examine how L1 subtitles and L2 captions influence primary EFL students' lexical development and content comprehension through viewing. The L1 subtitles group obtained significantly higher scores on the comprehension test than the L2 captions group, while the L2 captions group performed significantly better on a written word-form recognition task. In the form recognition test, the learners were presented with a set of L2 words and asked to circle only those words that appeared in the video. However, this test format may have disadvantaged the L1 subtitles group, who had not been exposed to these written word forms when they watched the videos (e.g., Jelani & Boers, 2018).

Bilingual Subtitles

Bilingual subtitles, also referred to as dual subtitles, are distinctive in that they display both L1 and L2 captions simultaneously on the screen. These can appear in two separate lines (vertical layout) or in two columns (horizontal layout) (see Figure 1). In Figure 1, panels a) and b) illustrate the horizontal layout, with L2 (English) on the left and L1 (Spanish) on the right. Panels c) and d) show the vertical layout, where L1 appears on the upper line and L2 on the lower line. Additionally, panels b) and c) include a visual separator (e.g., dots) between the L1 and L2 text to distinguish the two languages more clearly.

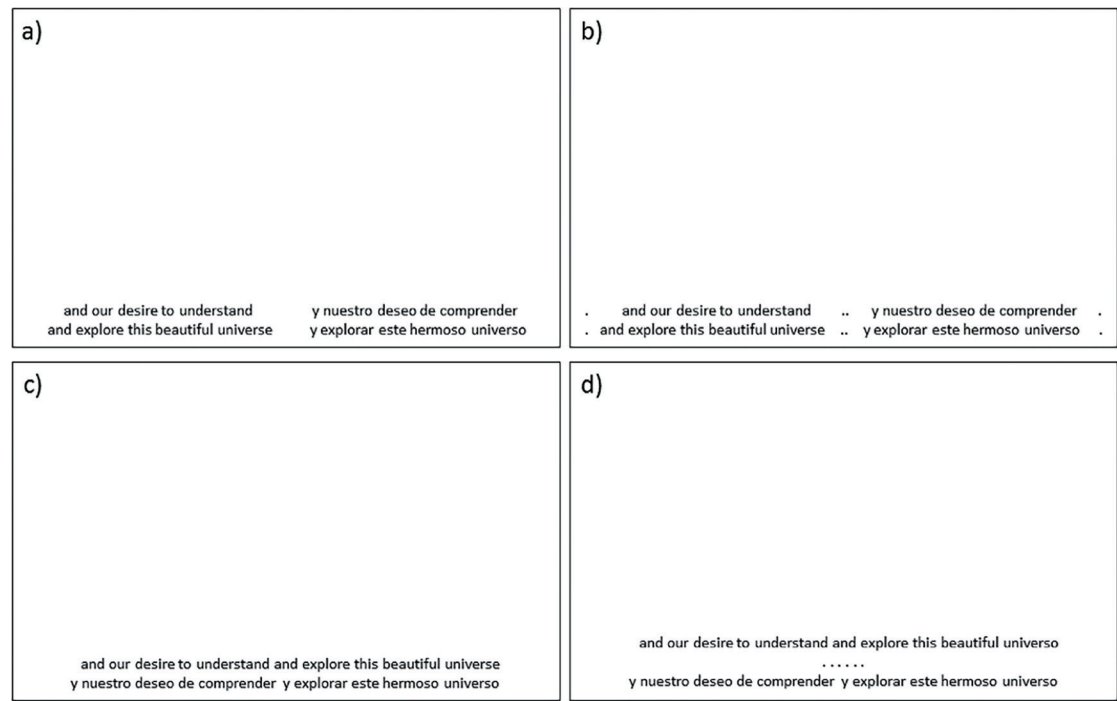


Figure 1. *Layouts of bilingual subtitles (García, 2017, p. 472)*

Among various formats of bilingual subtitles, the vertical layout without separators is the most commonly used. As a relatively new form of on-screen text, bilingual subtitles have gained popularity, especially in short videos on social media and YouTube videos, where bilingual creators use them to reach diverse audiences. Subtitles that include English are particularly effective for attracting global viewers. Advances in artificial intelligence have also made AI-powered auto subtitle generators more accessible, allowing content creators to produce bilingual subtitles with greater ease and accuracy (e.g., Ma, 2024). Reflecting this growing trend, researchers in Applied Linguistics and SLA have begun to examine their potential for language learning. It is hypothesized that bilingual subtitles may be especially beneficial for L2 vocabulary acquisition, as they provide simultaneous access to both the written form and the meaning of unfamiliar words (García, 2017; Li & Hennebry-Leung, 2022; Wi & Boers, 2014). However, research on bilingual subtitles remains limited compared to studies on monolingual captions and subtitles (e.g., García, 2017; Li & Hennebry-Leung, 2022; Lo, 2022; Teng, 2025; Wang & Pellicer-Sánchez, 2022).

García (2017) surveyed users of DualSub, an open-source tool for creating bilingual subtitles, to explore their perceptions of its usefulness for L2 vocabulary development. Most respondents were L2 learners (mainly English and German), with 77% reporting a preference for L2 captions while watching videos. However, 69.35% also reported that they switch to L1 captions when they encounter unfamiliar words, suggesting learners use the two types of on-screen text for different purposes. In an eye-tracking study, Wang and Pellicer-Sánchez (2022) compared the effects of bilingual subtitles, L2 captions, L1 subtitles, and no on-screen text on vocabulary learning among Chinese learners of English. Bilingual subtitles led to better performance than

L2 captions on meaning-recognition tests and better performance than L1 subtitles on meaning-recall tests, but were less effective than L2 captions for form recognition. Eye-tracking data showed that these learners attended most to the L1 text in the bilingual condition, contrasting with García's (2017) survey results and thus possibly indicating a difference between learners' self-reported use of bilingual subtitles and their actual processing of such materials. Li and Hennebry-Leung (2022) also found bilingual subtitles more effective than monolingual L2 or L1 captions for meaning recall and recognition. However, their study did not assess acquisition of word forms. Teng (2025) investigated the effect of bilingual subtitles on language minority students' learning of English multiword expressions in China. The results showed that bilingual subtitles facilitated meaning recognition; however, no evidence emerged of learning at the level of form and meaning recall.

While bilingual subtitles can support L2 vocabulary learning to some extent, processing both L1 and L2 captions simultaneously appears to be cognitively demanding due to learners' limited attentional capacity. Eye-tracking data suggest that the benefits of L2 captions may be compromised as learners focus mostly on the L1 lines (Wang & Pellicer-Sánchez, 2022). To address this issue, Wi and Boers (2024) proposed a sequential viewing method, in which L1 and L2 captions are presented in separate viewings, allowing learners to take full advantage of the benefits each type of caption offers. Learners first watch with L1 captions for comprehension, then with L2 captions to focus on L2 forms, or vice versa. This repeated viewing approach was tested in a mixed-methods study comparing an L1–L2–no captions sequence with an L2-only sequence, using a 5-minute TED Talk. The former sequence led to better results on a meaning-recall test, but the difference was not statistically significant, possibly owing to the sample size. Interestingly, access to L1 captions led to better performance on a quiz about the content of the TED Talk and this, in turn, predicted vocabulary gains. This illustrates how L1 captions can scaffold comprehension and how improved comprehension supports language learning. Some might question whether repeated viewing is a realistic approach. However, 94% of the participants in the study indicated that they regularly watch the same videos more than once. The reasons they mentioned for doing this included simply enjoying the content again, learning new words and expressions, and improving pronunciation. Repeated viewing has also been a feature of some other studies we reviewed earlier (e.g., Lo, 2024; Majuddin et al., 2021, 2024; Muñoz et al., 2024; Popova & Miralpeix, 2024).

Learner-related and Word-related Factors Affecting Vocabulary Learning through Audiovisual Input

Various factors that can influence vocabulary learning through audiovisual input have been examined, including learner-related, word-related, video-related (e.g., video length and genres), and methodological variables (e.g., Kurokawa et al., 2024). Among these, learner-related and word-related factors are the most frequently investigated. Therefore, this section focuses on these two categories of variables.

Learner-related factors found to be associated with greater vocabulary gains in the context of viewing include L2 proficiency level (e.g., Chen et al., 2018; Montero Perez et al., 2013; Muñoz et al., 2023; Teng, 2022), prior vocabulary knowledge as measured by vocabulary size or levels tests (e.g., Fievez et al., 2023; Montero Perez, 2020; Montero Perez et al., 2014, 2018; Peters et al., 2016; Peters & Webb, 2018; Puimège & Peters, 2019; Suárez & Gesa, 2019; Teng, 2023a; Teng & Mizumoto, 2023; Teng & Cui, 2025), working memory (e.g., Montero Perez, 2020; Teng, 2023a, 2023b; Teng & Cui, 2025), and aptitude for language learning (e.g., Casulleras & Miralpeix (2024); Muñoz et al., 2024; Suárez & Gesa, 2019; Teng, 2022). Muñoz et al. (2023) found that learners' proficiency levels significantly influenced L2 vocabulary learning from audiovisual input with on-screen text. Their study examined

three groups of learners at different proficiency levels, according to the Common European Framework of Reference (CEFR): Pre-A, A1, and A2/B1. The results showed that learners at the A2/B1 level outperformed those in the lower proficiency groups on both written form recall (prompted by the spoken forms) and meaning recall tasks, suggesting that higher proficiency is associated with greater lexical gains. Fievez et al. (2023) examined the effects of glossed captions on Dutch-speaking learners' incidental vocabulary learning in French, while also exploring the role of learners' prior vocabulary knowledge. The results revealed a positive association between vocabulary size and both form and meaning recall, indicating that learners with higher vocabulary size scores were more likely to perform better on the posttest. Teng (2023a) also found that learners' breadth of vocabulary knowledge, measured by the Vocabulary Levels Test (Schmitt et al., 2001), was positively associated with incidental vocabulary learning from captioned video at the levels of meaning recognition and recall. However, the association with depth of vocabulary knowledge (measured by means of Read's [1998] depth of vocabulary knowledge test) was much weaker. Conversely, however, Teng and Cui (2025) found that, while depth of vocabulary knowledge was a strong predictor of L2 collocation learning, breadth of vocabulary knowledge was not.

One of the emerging factors being explored in research on language learning from audiovisual input is working memory (WM). WM is the ability to temporarily hold and process information in short-term memory, supporting cognitive tasks such as comprehension, learning, and reasoning. Montero Perez (2020) investigated the role of working memory in vocabulary learning through video and found that complex working memory, as measured by a backward digit-span task and an operation span (OSPAN) task, significantly predicted vocabulary learning outcomes. Learners with higher scores on these complex working memory tasks were more likely to recognize the form and meaning of new words encountered in the video. Teng (2023a) found that complex working memory predicted performance on delayed meaning recognition and recall tests.

Another learner-related factor is language learning aptitude, which refers to an individual's ability or talent for acquiring a foreign or second language. Muñoz et al. (2024) examined the role of aptitude in vocabulary learning through captioned video. Aptitude was assessed using a sound recognition test and a reading efficacy test. The findings revealed that sound recognition skills significantly predicted performance on the meaning recognition test, whereas reading efficacy was not a strong predictor of vocabulary learning. A similar result was reported by Suárez and Gesa (2019), where aptitude was found to predict outcomes on a meaning recall test. However, research on the role of WM and aptitude in vocabulary learning through audiovisual input remains limited.

Word-related variables that have been investigated encompass frequency of occurrence in the material used in the particular study (e.g., Fievez et al., 2020, 2023; Muñoz et al., 2023; Peters, 2019; Peters et al., 2016; Rodgers & Webb, 2020), general corpus frequency (e.g., Puimège & Peters, 2019), and cognateness (e.g., Fievez et al., 2020; Peters & Webb, 2018). These factors are generally associated with enhanced L2 lexical development, although the effect might not always be statistically significant (e.g., Teng, 2022). Frequency of occurrence has been shown to predict L2 vocabulary learning through various modes of input, including reading (e.g., Pellicer-Sánchez, 2016), listening (e.g., Vidal, 2011), and viewing (e.g., Peters & Webb, 2018). In the context of viewing, Fievez et al. (2023) found that the frequency of occurrence of items significantly predicted performance on both form and meaning recall tests, aligning with previous findings by Rodgers and Webb (2020). Notably, they observed that the effect of frequency of occurrence is more pronounced when target words are encountered repeatedly within a single video, rather than across multiple videos.

Corpus frequency of words has also been found to affect L2 lexical development through viewing. Puimège and Peters (2019) investigated how word-related factors such as corpus frequency, number of syllables, part of speech, and concreteness affect the learning of a single word and formulaic sequences (FS) through viewing. The results revealed that corpus frequency was a significant predictor of both single-word and formulaic sequence (FS) learning at the level of form recall. They also found that concreteness and word length were positively correlated with single-word learning. Concreteness emerged as the most significant predictor, yielding the highest odds ratio ($Exp(B) = 5.03$), confirming earlier research (e.g., de Groot & Keijzer, 2000) that concrete/imageable words are more likely to be learned than abstract ones (which is in keeping with Paivio's Dual Coding Theory, mentioned earlier). It is plausible that the role of concreteness is particularly pronounced in the case of audiovisual input, because it is concrete referents that are most readily depicted or illustrated in such materials. However, it should be noted that Puimège and Peters' (2019) study was conducted with a small sample size, underscoring the need for further research to validate the effects of these word-related factors (e.g., corpus frequency, word length, concreteness) specifically in the context of viewing.

Another word-related factor that is known to influence the pace of acquisition is cognateness (e.g., Lotto & de Groot, 1998). Cognates are L1–L2 words that are closely related in their phonological or orthographic form, meaning, and origin. A recent study by Fievez et al. (2020) demonstrated a significant positive relationship between cognates and L2 word learning through captioned video. The study, which involved Dutch-speaking learners of French (i.e., a language pair with a considerable number of cognates), revealed that the meanings of cognates were 8% and 20% more likely to be recognized and recalled, respectively, compared to non-cognates. These findings align with earlier evidence by Peters and Webb (2018).

A Recurring Methodological Challenge: Test Modality and Formats

A challenge for researchers in this strand of inquiry (and in studies of vocabulary learning more generally) is to choose among possible measures of vocabulary learning and interpret participants' performance on the chosen measures. For example, students who have watched a video without L2 captions may perform poorly on a test that uses written word prompts simply because they have not encountered the written forms of these words. Presented with the spoken forms, they may stand a better chance of recognizing them and possibly recalling what concepts they related to in the video. This matter of congruence in modality need not be an issue in the case of languages with a consistent correspondence between spoken and written forms (where the spoken form of a word is a reliable predictor of what its written form looks like, and vice versa). However, it can definitely be problematic in the case of a language such as English. For example, Jelani and Boers (2018) compared the effects of captioned versus uncaptioned videos on intermediate learners' vocabulary uptake and found a positive effect of captions only when the test prompts were written. No advantage for captions was observed when the prompts were presented aurally, indicating that test format can play a mediating role in learners' performance. Such a mediating role was also observed in an earlier study, by Winke et al. (2010), who investigated whether captioning during the first or second viewing was more beneficial. Native English speakers learning Spanish, Russian, Arabic, or Chinese watched a video and translated target words into English. Half of the items were presented aurally and half in written form. Learners who first watched the captioned version performed the best on this test, but this held true particularly for the aurally presented items. Seeing the captioned version first may have helped the learners to identify the L2 words in the audio-recording when they subsequently watched the uncaptioned version. In any case, it is reassuring to see that many recent studies have taken the question of congruence of input modality and test modality into consideration

by either presenting some prompts in written form and others in spoken form or by presenting all items in both modalities.

In studies that aim to compare vocabulary learning from viewing videos with L2 captions, L1 subtitles, or dual subtitles (e.g., Li & Hennebry-Leung, 2022; Wang & Pellicer-Sánchez, 2022), it may matter if the test prompts are presented in L1 or L2. Participants who have not seen or have not attended to L2 captions may find it hard to recognize the L2 words in a post-test (again, particularly in a language such as English, where recalling the spoken form of words does not always help to predict their written form).

The choice of test format may, of course, matter as well. L2 captions combined with contextual cues can help learners to interpret the meanings of new words, but these interpretations will often be imprecise. In comparison, L1 subtitles directly provide the correct meanings and thus leave no ambiguity. A meaning recognition test (typically a multiple-choice test) may enable learners who have watched a video with L2 captions to select the correct candidate responses if they match the learners' hunches. Such a test may then yield scores that are similar for learners who have watched the video with L2 captions and those who have watched it with L1 subtitles, while the latter on-screen text may, in fact, have been more beneficial to establish word meanings. It may require a (more challenging) meaning-recall test (i.e., where the learners themselves need to supply the meaning of a word) to determine if watching the video with L1 subtitles promotes more precise semantic knowledge.

Limitations of the Available Research

This review of the literature reveals several areas in the research on L2 vocabulary learning from audiovisual input that invite further exploration. One is that there is very little research about the role of the nature of the audiovisual input. Studies have investigated the affordances of various types of captions when applied to TED Talks, TV shows, and documentaries. However, they have not investigated whether captioning might be more useful (or necessary) for certain types or genres than others (e.g., genres with fast or slow speech rates, or genres with varying amounts of visual-contextual support). For instance, learners watching a wildlife documentary where a slow narrative simply describes events shown on the screen may not need the support of captions as much as when they watch a fast-paced televised debate without many visual cues to help interpretation. We are aware of only one study (Teng, 2023a) that has compared the benefits of captioning when applied to different video genres (comedy, education, and documentary). The findings revealed that vocabulary learning at the meaning recall level occurred across all genres, with the highest learning gains observed in the participant group that viewed the comedy genre. A challenge with this novel research endeavor is that it is hard to isolate the role of genre as such, unless various other variables that influence learning gains, such as topic familiarity and the characteristics of the target words, are taken into consideration.

It is noteworthy that research on the benefits of L1 subtitles remains limited compared to research on L2 captions. L2 captions have been widely studied and shown to be effective for learning word forms. However, they are often insufficient for developing semantic knowledge, as “captioning does not provide explicit access to word meaning” (Montero Perez et al., 2018, p. 3). Contextual clues and imagery may help form-meaning mappings, but they “do not always provide explicit word meaning information” (Montero Perez, 2022, p. 173). Conversely, L1 subtitles offer direct access to meaning, but they are less helpful for learning the form of L2 words. They are naturally less helpful for learning the written L2 forms as these are absent from the L1 subtitles, but they are also less helpful for learning the spoken L2 forms because subtitles are no transcript of the L2 speech. L1 subtitles and L2 captions thus complement one another.

It stands to reason that the combined use of L1 and L2 captions promotes lexical development, as both form and meaning are essential facets of word knowledge (Nation, 2022).

It is true that glossed keyword captions also provide lexical meanings and could therefore be an alternative for subtitles. However, glossed keyword captions are uncommon, and the technological demands are high should teachers themselves wish to create them for use with their students. A more common way of providing both written form and meaning is through dual subtitles. However, owing to the real-time nature of video viewing, learners struggle to read both lines of text, as evidenced by eye-tracking studies. Even if they were able to read all of the on-screen text before it is replaced by subsequent text, the L1 text may make the L2 text redundant for the purpose of understanding the video content. Moreover, despite their increasing use (e.g., DualSub, Language Reactor), dual subtitles are still less commonly available than monolingual captions, limiting their broader applicability in language learning settings. A potential solution is the sequential use of L1 and L2 captions, as proposed in an exploratory study by Wi and Boers (2024). However, to our knowledge, no empirical studies have directly compared the benefits of dual subtitles and the benefits of the sequential use of L1 and L2 captions for vocabulary learning.

Additionally, most existing studies on vocabulary learning through viewing are entirely input-based: Participants are asked to watch a video and are subsequently given a quiz about its content and the lexical items that were used in the video. In language courses as well as outside the language classroom, it is likely that the content of newsworthy, engaging, or thought-provoking videos prompts viewers to communicate about it (such as telling someone else about it or exchanging reactions to it). When learners use their L2 to express themselves, this creates opportunities for experimenting with new words or phrases, solidifying knowledge of newly acquired items by retrieving them from memory, and noticing lacunae in their lexical repertoires, which is all in keeping with the Output Hypothesis (Swain, 1995). Nguyen and Boers (2019) demonstrated that asking students to talk about a TED Talk they had just watched, and then having them watch the same Talk once more, improved the likelihood of vocabulary learning, compared to a condition without the output activity. However, their study of this input-output-input cycle used a video without on-screen text. A promising avenue could be to incorporate on-screen text into such cycles to reap the combined benefits of captions and output activities.

Finally, much of the existing research in this field has been conducted under so-called incidental learning conditions, where learners are expected to focus on the content of input materials rather than the wording of this content. While some studies clearly identified their design as focusing on incidental vocabulary learning, others have been labelled as such by default simply because participants were not told in advance about upcoming vocabulary tests (see, e.g., Webb, 2020, for discussion). However, not informing learners about a vocabulary post-test does not rule out the possibility that they intentionally engage with new words or phrases, especially if the study is conducted in a language course or when the participants are first given a pre-test. Although there is no denying that watching videos in the target language can contribute significantly to vocabulary development, it is uncertain whether the learning gains reported in this strand of research are reflective of the amount of learning that occurs as a side-benefit of watching videos for pleasure outside the language classroom or outside a setting where learners have volunteered to participate in a study. It stands to reason that L2 learners at least occasionally shift their attention from content to language when they notice a new lexical item that piques their curiosity and/or that they recognize is essential to follow the content. Arguably, without this “noticing,” learning will not happen in the first place (Schmidt, 1990, 1993). We could define incidental language acquisition as “the acquisition of items or patterns of language over the course of an activity where the learner is primarily engaged with the content of a message,

not its linguistic packaging” (Boers, 2021, p. 17), where inclusion of “primarily” in the definition recognizes that it is hard in actual practice to gauge the extent to which learners switch to a more intentional learning mode during activities that were intended to be content-focused. This adds to the many variables at play that collectively help to explain the variability in attested learning outcomes in this strand of research. It would be useful for future research to routinely include data on learner-participants’ perceptions of content-focused tasks and whether they were aware of (or suspected) the (true) purpose of the study they signed up for.

Conclusions, Directions, and Pedagogical Implications

Audiovisual materials as a source for L2 acquisition—especially vocabulary acquisition—have attracted considerable attention in the past two decades. Research has demonstrated that vocabulary acquisition happens through engagement with various types of multimodal input, provided that it is sufficiently comprehensible to the learners. A substantial collection of studies has also demonstrated that the rate of vocabulary acquisition increases when learners are offered support in the form of on-screen text. Some studies in this line of research have investigated how repeated viewing of the same material mediates the benefits of onscreen text. Other studies investigated whether the onscreen text can be made more helpful by making certain lexical items visually salient or by providing word meanings in the form of glosses. Recently, the affordances of bilingual subtitles have started to attract researchers’ attention as well.

Future research could further explore how on-screen text can be used to promote learning of both the (spoken and written) forms of L2 words and their meanings. For instance, no studies have yet compared the effects of bilingual subtitles and the sequential use of L1 and L2 captions. Considering their potential facilitative effects on L2 vocabulary acquisition and the increasing advances in technology, comparing these two methods could offer valuable insights into effective and efficient strategies for language learning through audiovisual input. To gain a deeper understanding of learners’ experiences with these viewing methods, future studies should adopt a triangulated methodology, incorporating not only measures of vocabulary learning, but additional data from eye-tracking, questionnaires and/or retrospective interviews.

Another promising research direction is the integration of output tasks (e.g., speaking or writing) based on the content of audiovisual input. Language learning is most effective when learners are engaged in both input and output processes (e.g., Nguyen & Boers, 2019), combining learning opportunities recognized by the Input Hypothesis (e.g., Krashen, 1982; Krashen & Terrell, 1983) and by the Output Hypothesis (e.g., Swain, 1995). However, most existing studies on L2 lexical learning through video to date have focused solely on the effects of input. Research examining the impact of input-output cycles may inform pedagogy by helping teachers implement a more balanced approach across the four language skills (listening, speaking, reading, and writing).

Future research on incidental vocabulary learning through video should also incorporate triangulated methods to ascertain that reported learning outcomes genuinely reflect incidental learning. Follow-up questionnaires simply asking whether learners expected a vocabulary test, or open-ended questions allowing learners to describe how they interacted with the video and onscreen text, can shed light on their level of intentionality. Additionally, more research is needed on intentional vocabulary learning through audiovisual input, especially with the use of L1 captions, L2 captions, and bilingual subtitles as scaffolding tools. These forms of textual support can be used strategically and purposefully by learners to enhance vocabulary acquisition and overall language development (e.g., sequencing different caption types across repeated viewings or manipulating the presence of L1 and L2 captions). Vanderplank and

Teng (2024) suggest that learners should engage in captioned viewing consciously “to gain more tangible benefits” (p. 21). Findings from studies on intentional vocabulary learning can also provide evidence-based strategies that learners and educators can adopt to make more informed and effective use of audiovisual materials for L2 lexical development in instructed second language acquisition contexts.

References

- Alobaid, A. (2021). ICT multimedia learning affordances: role and impact on ESL learners’ writing accuracy development. *Heliyon*, 7(7), e07517–e07517. <https://doi.org/10.1016/j.heliyon.2021.e07517>
- Ariew, R., & Erçetin, G. (2004). Exploring the potential of hypermedia annotations for second language reading. *Computer Assisted Language Learning*, 17, 237–259.
- Baltova, I. (1994). The impact of video on the comprehension skills of core French students. *Canadian Modern Language Review*, 50(3), 507–531. <https://doi.org/10.3138/cmlr.50.3.507>
- Baranowska, K. (2020). Learning most with least effort: subtitles and cognitive load. *ELT Journal*, 74(2), 105–115. <https://doi.org/10.1093/elt/ccz060>
- Bird, S., & Williams, J. (2002). The effect of bimodal input on implicit and explicit memory: An investigation into the benefits of within-language subtitling. *Applied Psycholinguistics*, 23(4), 509–533. <https://doi.org/10.1017/S0142716402004022>
- Birulés-Muntané, J., & Soto-Faraco, S. (2016). Watching subtitled films can help learning foreign languages. *PLoS One*, 11(6), e0158409–e0158409. <https://doi.org/10.1371/journal.pone.0158409>
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.
- Bjork, R. A. (2018). Being suspicious of the sense of ease and undeterred by the sense of difficulty: Looking back at Schmidt and Bjork (1992). *Perspectives on Psychological Science*, 13(2), 146–148. <https://doi.org/10.1177/1745691617690642>
- Boers, F. (2013). Cognitive linguistic approaches to teaching vocabulary: Assessment and integration. *Language Teaching*, 46(2), 208–224. <https://doi.org/10.1017/S0261444811000450>
- Boers, F. (2021). *Evaluating second language vocabulary and grammar instruction: A synthesis of the research on teaching words, phrases, and patterns*. Routledge.
- Boers, F., Demecheleer, M., He, L., Deconinck, J., Stengers, H., & Eyckmans, J. (2017). Typographic enhancement of multiword units in second language text. *International Journal of Applied Linguistics*, 27(2), 448–469. <https://doi.org/10.1111/ijal.12141>
- Braun, V., & Clarke, V. (2022). Conceptual and design thinking for thematic analysis. *Qualitative Psychology*, 9(1), 3–26. <https://doi.org/10.1037/qup0000196>
- Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, 20, 136–163. <https://doi.org/10.125/66816>
- Carpenter, S. K., Sachs, R. E., Martin, B., Schmidt, K., & Looft, R. (2012). Learning new vocabulary in German: The effects of inferring word meanings, type of feedback, and time of test. *Psychonomic Bulletin & Review*, 19(1), 81–86. <https://doi.org/10.3758/s13423-011-0185-7>
- Casulleras, M., & Miralpeix, I. (2024). Language learning from watching cartoons in the primary EFL classroom. In C. Muñoz & I. Miralpeix (Eds.), *Audiovisual input and second language learning* (Vol. 61, pp. 1–24). John Benjamins Publishing Company. <https://doi.org/10.1075/llt.61.07mor>
- Charles, T. J., & Trenkic, D. (2015). Speech segmentation in a second language: The role of bimodal input. In Y. Gambier, A. Caimi, & C. Mariotti (Eds.), *Subtitles and language learning: Principles, strategies and practical experiences*, (pp. 173–198). Peter Lang.
- Chen, Y.-R., Liu, Y.-T., & Todd, A. G. (2018). Transient but effective? Captioning and adolescent EFL learners’ spoken vocabulary acquisition. *English Teaching and Learning*, 42(1), 25–56. <https://doi.org/10.1007/s42321-018-0002-8>
- Choi, S. (2017). Processing and learning of enhanced English collocations: An eye-movement study. *Language Teaching Research*, 21(3), 403–426. <https://doi.org/10.1177/1362168816653271>
- Cintrón-Valentín, M. C., & García-Amaya, L. (2021). Investigating textual enhancement and captions in L2 grammar and vocabulary: An experimental study. *Studies in Second Language Acquisition*, 43(5), 1068–1093. <https://doi.org/10.1017/S0272263120000492>

- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE Publications.
- Danan, M. (1992). Reversed subtitling and dual coding theory: New directions for foreign language instruction. *Language Learning*, 42(4), 497–527. <https://doi.org/10.1111/j.1467-1770.1992.tb01042.x>
- Danan, M. (2004). Captioning and subtitling: Undervalued language learning strategies. *Meta: Journal Des Traducteurs*, 49(1), 67–77. <https://doi.org/10.7202/009021ar>
- Dang, T. N. Y., Lu, C., & Webb, S. (2022). Incidental learning of single words and collocations through viewing an academic lecture. *Studies in Second Language Acquisition*, 44(3), 708–736. <https://doi.org/10.1017/S0272263121000474>
- de Groot, A. M. B., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign language vocabulary learning and forgetting. *Language Learning*, 50(1), 1–56. <https://doi.org/10.1111/0023-8333.00110>
- d'Ydewalle, G., & Pavakanun, U. (1997). Could enjoying a movie lead to language acquisition? In P. Winterhoff-Spurk & T. van der Voort (Eds.), *New horizons in media psychology* (pp. 145–155). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-663-10899-3_10
- Elgort, I., Beliaeva, N., & Boers, F. (2020). Contextual word learning in the first and second language: Definition placement and inference error effects on declarative and nondeclarative knowledge. *Studies in Second Language Acquisition*, 42(1), 7–32. <https://doi.org/10.1017/S0272263119000561>
- Elgort, I. (2017). Incorrect inferences and contextual word learning in English as a second language. *Journal of the European Second Language Association*, 1(1), 1–11. <https://doi.org/10.22599/jesla.3>
- Farley, A. P., Ramonda, K., & Liu, X. (2012). The concreteness effect and the bilingual lexicon: The impact of visual stimuli attachment on meaning recall of abstract L2 words. *Language Teaching Research*, 16(4), 449–466. <https://doi.org/10.1177/1362168812436910>
- Fievez, I., Perez, M. M., Cornillie, F., & Desmet, P. (2020). Vocabulary learning through viewing captioned or subtitled videos and the role of learner-and word-related factors. *CALICO Journal*, 37(3), 21–41. <https://doi.org/10.1558/cj.39370>
- Fievez, I., Montero Perez, M., Cornillie, F., & Desmet, P. (2023). Promoting incidental vocabulary learning through watching a French Netflix series with glossed captions. *Computer Assisted Language Learning*, 36(1–2), 26–51. <https://doi.org/10.1080/09588221.2021.1899244>
- Finger-Bou, R., & Muñoz, C. (2024). The effects of textual enhancement on young learners' attention and vocabulary acquisition through captioned cartoons. In C. Muñoz & I. Miralpeix (Eds.), *Audiovisual input and second language learning* (pp. 48–73). John Benjamins Publishing Company. <https://doi.org/10.1075/llt.61.03fin>
- Fraser, C. A. (1999). Lexical processing strategy use and vocabulary learning through reading. *Studies in Second Language Acquisition*, 21(2), 225–241. <https://doi.org/10.1017/S0272263199002041>
- Frumuselu, A., De Maeyer, S., Donche, V., & Colon Plana, M. (2015). Television series inside the EFL classroom: Bridging the gap between teaching and learning informal language through subtitles. *Linguistics and Education*, 32, 107–117. <https://doi.org/10.1016/j.linged.2015.10.001>
- García, B. (2017). Bilingual subtitles for second-language acquisition and application to engineering education as learning pills. *Computer Applications in Engineering Education*, 25(3), 468–479. <https://doi.org/10.1002/cae.21814>
- Han, Z., Park, E. S., & Combs, C. (2008). Textual enhancement of input: Issues and possibilities. *Applied Linguistics*, 29, 597–618. <https://doi.org/10.1093/applin/amn010>
- Hao, T., Sheng, H., Ardasheva, Y., & Wang, Z. (2022). Effects of dual subtitles on Chinese students' English listening comprehension and vocabulary learning. *The Asia-Pacific Education Researcher*, 31(5), 529–540. <https://doi.org/10.1007/s40299-021-00601-w>
- Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a foreign language*, 11(2), 207–223. <https://doi.org/10.125/66953>
- Hsieh, Y. (2020). Effects of video captioning on EFL vocabulary learning and listening comprehension. *Computer Assisted Language Learning*, 33(5–6), 567–589. <https://doi.org/10.1080/09588221.2019.1577898>
- Jelani, N. A. M., & Boers, F. (2018). Examining incidental vocabulary acquisition from captioned video. *ITL-International Journal of Applied Linguistics*, 169(1), 169–190. <https://doi.org/10.1075/itl.00011.jel>
- Kang, E. (2019). Enhancing Korean EFL learners' vocabulary learning and listening comprehension through video captions. *STEM Journal*, 20(2), 91–108. <https://doi.org/10.16875/stem.2019.20.2.91>

- Koolstra, C., & Beentjes, J. (1999). Children's vocabulary acquisition in a foreign language through watching subtitled television programs at home. *Educational Technology Research and Development*, 47(1), 51–60. <https://doi.org/10.1007/BF02299476>
- Krashen, S. (1982). *Principles and practice in second language acquisition*. Pergamon Press.
- Krashen, S. D., & Terrell, T. D. (1983). *The natural approach: Language acquisition in the classroom* (1st ed.). Pergamon Press.
- Kruger, J.-L., Hefer, E., & Matthew, G. (2014). Attention distribution and cognitive load in a subtitled academic lecture: L1 vs. L2. *Journal of Eye Movement Research*, 7(5). <https://doi.org/10.16910/jemr.7.5.4>
- Kuppens, A. (2010). Incidental foreign language acquisition from media exposure. *Learning, Media and Technology*, 35(1), 65–85. <https://doi.org/10.1080/17439880903561876>
- Kurokawa, S., Hein, A. M., & Uchihara, T. (2024). Incidental vocabulary acquisition through captioned viewing: A meta-analysis. *Language Learning*, 1–49. <https://doi.org/10.1111/lang.12697>
- Lee, T., Lee, Y., & Choi, S. (2025). First- and second-language subtitles and cognitive load: An EEG study. *Language Learning & Technology*, 29(1), 1–18. <https://hdl.handle.net/10125/73625>
- Lee, H., & Mayer, R. E. (2015). Visual aids to learning in a second language: Adding redundant video to an audio lecture. *Applied Cognitive Psychology*, 29(3), 445–454. <https://doi.org/10.1002/acp.3123>
- Lee, H., & Mayer, R. E. (2018). Fostering learning from instructional video in a second language. *Applied Cognitive Psychology*, 32(5), 648–654. <https://doi.org/10.1002/acp.3436>
- Lee, M., & Révész, A. (2018). Promoting grammatical development through textually enhanced captions: An eye-tracking study. *The Modern Language Journal*, 102(3), 557–577. <https://doi.org/10.1111/modl.12503>
- Lee, M., & Révész, A. (2020). Promoting grammatical development through captions and textual enhancement in multimodal input-based tasks. *Studies in Second Language Acquisition*, 42(3), 625–651. <https://doi.org/10.1017/S0272263120000108>
- Li, M., & Hennebry-Leung, M. (2024). Effects of monolingual and bilingual subtitles on L2 vocabulary acquisition. *International Review of Applied Linguistics in Language Teaching*, 62(2), 843–870. <https://doi.org/10.1515/iral-2022-0034>
- Lo, S. (2022). Learning vocabulary through dual-subtitled viewing: The impact of different ILH-based interventions. *Computer Assisted Language Learning*, 1–28. <https://doi.org/10.1080/09588221.2022.2126497>
- Lo, S. (2024). Vocabulary learning through viewing dual-subtitled videos: Immediate repetition versus spaced repetition as an enhancement strategy. *ReCALL*, 36(2), 152–167. <https://doi.org/10.1017/S0958344024000053>
- Lotto, L., & de Groot, A. M. B. (2002). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning*, 48(1), 31–69. <https://doi.org/10.1111/1467-9922.00032>
- Lwo, L., & Lin, M. C. -T. (2012). The effects of captions in teenagers' multimedia L2 learning. *ReCAL*, 24(2), 188–208. <https://doi.org/10.1017/S0958344012000067>
- Ma, A. (2024). AI vs. human translation: Shaping the future of language learning through AI-generated bilingual captions. *SHS Web of Conferences*, 199, 01023. <https://doi.org/10.1051/shsconf/202419901023>
- Majuddin, E., Boers, F., & Siyanova-Chanturia, A. (2024). The effects of enhancing L2 multiword items in captions: An approximate replication of Majuddin, Siyanova-Chanturia, and Boers (2021). *Language Teaching*, 1–18. <https://doi.org/10.1017/S0261444824000296>
- Majuddin, E., Siyanova-Chanturia, A., & Boers, F. (2021). Incidental acquisition of multiword expressions through audiovisual materials: The role of repetition and typographic enhancement. *Studies in Second Language Acquisition*, 43(5), 985–1008. <https://doi.org/10.1017/S0272263121000036>
- Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist*, 32(1), 1–19. https://doi.org/10.1207/s15326985sep3201_1
- Mayer, R. E. (2011). *Applying the science of learning*. Pearson.
- Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 43–71). Cambridge University Press.
- Mayer, R. E. (2017). Using multimedia for e-learning. *Journal of Computer Assisted Learning*, 33(5), 403–423. <https://doi.org/10.1111/jcal.12197>
- Mayer, R. E. (2021). *Multimedia learning* (3rd ed.). Cambridge University Press.

- Mayer, R. E., & Anderson, R. W. (1992). The instructive animation: Helping students build connections between words and pictures in multimedia learning. *Journal of Educational Psychology*, 84, 444–452. <https://doi.org/10.1037/0022-0663.84.4.444>
- Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand words? *Journal of Educational Psychology*, 82, 715–726. <https://doi.org/10.1037/0022-0663.82.4.715>
- Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6
- Mitterer, H., & McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PLoS One*, 4(11), e7785–e7785. <https://doi.org/10.1371/journal.pone.0007785>
- Mondria, J.-A. (2003). The effects of inferring, verifying, and memorizing on the retention of L2 word meanings: An experimental comparison of the “meaning-inferred Method” and the “meaning-given method”. *Studies in Second Language Acquisition*, 25(4), 473–499. <https://doi.org/10.1017/S0272263103000202>
- Montero Perez, M., Peters, E., & Desmet, P. (2014a). Is less more? Effectiveness and perceived usefulness of keyword and full captioned video for L2 listening comprehension. *ReCALL*, 26(1), 21–43. <https://doi.org/10.1017/S0958344013000256>
- Montero Perez, M., Peters, E., & Desmet, P. (2015). Enhancing vocabulary learning through captioned video: An eye-tracking study. *The Modern Language Journal*, 99(2), 308–328. <https://doi.org/10.1111/modl.12215>
- Montero Perez, M., Peters, E., & Desmet, P. (2018). Vocabulary learning through viewing video: the effect of two enhancement techniques. *Computer Assisted Language Learning*, 31(1–2), 1–26. <https://doi.org/10.1080/09588221.2017.1375960>
- Montero Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014b). Effects of captioning on video comprehension and incidental vocabulary learning. *Language Learning & Technology*, 18(1), 118–141. <http://dx.doi.org/10.125/44357>
- Montero Perez, M., Van Den Noortgate, W., & Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System*, 41(3), 720–739. <https://doi.org/10.1016/j.system.2013.07.013>
- Montero Perez, M. (2020). Incidental vocabulary learning through viewing video: The role of vocabulary knowledge and working memory. *Studies in Second Language Acquisition*, 42(4), 749–773. <https://doi.org/10.1017/S0272263119000706>
- Montero Perez, M. (2022). Second or foreign language learning through watching audio-visual input and the role of on-screen text. *Language Teaching*, 55(2), 163–192. <https://doi.org/10.1017/S0261444821000501>
- Mora, J. C., & Fouz-González, J. (2024). Contrastive input enhancement in captioned video for L2 pronunciation learning. In C. Muñoz & I. Miralpeix (Eds.), *Audiovisual input and second language learning* (Vol. 61, pp. 150–175). John Benjamins Publishing Company. <https://doi.org/10.1075/llt.61.07mor>
- Moreno, R., & Mayer, R. (2007). Interactive multimodal learning Environments. *Educational Psychology Review*, 19(3), 309–326. <https://doi.org/10.1007/s10648-007-9047-2>
- Muñoz, C., Pujadas, G., & Pattemore, A. (2023). Audio-visual input for learning L2 vocabulary and grammatical constructions. *Second Language Research*, 39(1), 13–37. <https://doi.org/10.1177/02676583211015797>
- Muñoz, C., Pattemore, A., & Avello, D. (2024). Exploring repeated captioning viewing as a way to promote vocabulary learning: Time lag between repetitions and learner factors. *Computer Assisted Language Learning*, 37(7), 1744–1770. <https://doi.org/10.1080/09588221.2022.2113898>
- Nation, I. S. P. (2022). *Learning vocabulary in another language* (3rd ed.). Cambridge University Press.
- Neuman, S. B., & Koskinen, P. (1992). Captioned television as comprehensible input: Effects of incidental word learning from context for language minority students. *Reading Research Quarterly*, 27, 94–106.
- Nguyen, C., & Boers, F. (2019). The effect of content retelling on vocabulary uptake from a TED Talk. *TESOL Quarterly*, 53(1), 5–29. <https://doi.org/10.1002/tesq.441>
- OECD (2009). Special Focus: Measuring Leisure in OECD Countries. *Society at a glance 2009: OECD social indicators* (pp. 19–49). OECD Publishing. https://doi.org/10.1787/soc_glance-2008-3-en
- Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195066661.001.0001>

- Paivio, A. (2007). *Mind and its evolution: A dual coding theoretical approach*. L. Erlbaum Associates.
- Paivio, A. (2010). Dual coding theory and the mental lexicon. *The Mental Lexicon*, 5, 205–230.
- Pattemore, A., & Muñoz, C. (2020). Learning L2 constructions from captioned audio-visual exposure: The effect of learner-related factors. *System*, 93, 102303. <https://doi.org/10.1016/j.system.2020.102303>
- Pattemore, A., Suárez, M. del M., Montero Perez, M., & Muñoz, C. (2024). The role of language aptitude in learning L2 constructions from captioned and uncaptioned audiovisual input. In C. Muñoz & I. Miralpeix (Eds.), *Audiovisual input and second language learning* (Vol. 61, pp. 176–198). John Benjamins Publishing Company. <https://doi.org/10.1075/llt.61.08pat>
- Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do things fall apart? *Reading in a Foreign Language*, 22, 31–55. <https://doi.org/10.125/66652>
- Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition from and while reading. *Studies in Second Language Acquisition*, 38, 97–130. <https://doi.org/10.1017/S0272263115000224>
- Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition*, 40(3), 551–577. <https://doi.org/10.1017/S0272263117000407>
- Peters, E., Heynen, E., & Puimège, E. (2016). Learning vocabulary through audiovisual input: The differential effect of L1 subtitles and captions. *System*, 63, 134–148. <https://doi.org/10.1016/j.system.2016.10.002>
- Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *TESOL Quarterly*, 53(4), 1008–1032. <https://doi.org/10.1002/tesq.531>
- Popova, M., & Miralpeix, I. (2024). Maximizing L2 learning from captioned TV viewing. In C. Muñoz & I. Miralpeix (Eds.), *Audiovisual input and second language learning* (Vol. 61, pp. 100–125). John Benjamins Publishing Company. <https://doi.org/10.1075/llt.61.05pop>
- Puimège, E., & Peters, E. (2019). Learning L2 vocabulary from audiovisual input: An exploratory study into incidental learning of single words and formulaic sequences. *The Language Learning Journal*, 47(4), 424–438. <https://doi.org/10.1080/09571736.2019.1638630>
- Puimège, E., Montero Perez, M., & Peters, E. (2023). Promoting L2 acquisition of multiword units through textually enhanced audiovisual input: An eye-tracking study. *Second Language Research*, 39(2), 471–492. <https://doi.org/10.1177/02676583211049741>
- Pujadas, G., & Muñoz, C. (2019). Extensive viewing of captioned and subtitled TV series: A study of L2 vocabulary learning by adolescents. *The Language Learning Journal*, 47(4), 479–496. <https://doi.org/10.1080/09571736.2019.1616806>
- Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41–60). Erlbaum.
- Rodgers, M. P. H. (2018). The images in television programs and the potential for learning unknown words: The relationship between on-screen imagery and vocabulary. *ITL-International Journal of Applied Linguistics*, 169(1), 191–211. <https://doi.org/10.1075/itl.00012.rod>
- Rodgers, M. P. H., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly*, 45, 689–717. <https://doi.org/10.5054/tq.2011.268062>
- Rodgers, M. P. H., & Webb, S. (2020). Incidental vocabulary learning through viewing television. *ITL-International Journal of Applied Linguistics*, 171(2), 191–220. <https://doi.org/10.1075/itl.18034.rod>
- Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129–158. <https://doi.org/10.1093/applin/11.2.129>
- Schmidt, R. (1993). Awareness and second language acquisition. *Annual Review of Applied Linguistics*, 13(1), 206–226. [https://doi.org/10.1017/j.1467-9280.1992.tb00029.x](https://doi.org/10.1017/info:doi/10.1017/j.1467-9280.1992.tb00029.x)
- Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3(4), 207–217. <https://doi.org/10.1111/j.1467-9280.1992.tb00029.x>
- Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18(1), 55–88. <https://doi.org/10.1177/026553220101800103>
- Shen, H. H. (2010). Imagery and verbal coding approaches in Chinese vocabulary instruction. *Language Teaching Research*, 14(4), 485–499. <https://doi.org/10.1177/1362168810375370>

- Statistics Canada (2019). Daily average time spent in hours on various activities by age group and sex, 15 years and over, Canada and provinces [Data set]. <https://doi.org/10.25318/4510001401-eng>
- Suárez, M. del M., & Gesa, F. (2019). Learning vocabulary with the support of sustained exposure to captioned video: do proficiency and aptitude make a difference? *Language Learning Journal*, 47(4), 497–517. <https://doi.org/10.1080/09571736.2019.1617768>
- Swain, M., Kinnear, P., & Steinman, L. (2015). *Sociocultural theory in second language education: An introduction through narratives* (2nd edition). Multilingual Matters.
- Swain, M. (1995). Three functions of output in second language learning. In G. Cook & G. Seidhofer (Eds.), *Principles and practices in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125–144). Oxford University Press.
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory* (1st ed., Vol. 1). Springer. <https://doi.org/10.1007/978-1-4419-8126-4>
- Sydorenko, T. (2010). Modality of input and vocabulary acquisition. *Language Learning & Technology*, 2(14), 50–73. <http://dx.doi.org/10125/44214>
- Teng, F. (2018). Incidental vocabulary acquisition from reading-only and reading-while listening: A multi-dimensional approach. *Innovation in Language Learning and Teaching*, 12(3), 274–288. <https://doi.org/10.1080/17501229.2016.1203328>
- Teng, M. F. (2022). Incidental L2 vocabulary learning from viewing captioned videos: Effects of learner-related factors. *System (Linköping)*, 105, 102736-. <https://doi.org/10.1016/j.system.2022.102736>
- Teng, M. F. (2023a). Incidental vocabulary learning from captioned video genres: vocabulary knowledge, comprehension, repetition, and working memory. *Computer Assisted Language Learning*, 1–40. <https://doi.org/10.1080/09588221.2023.2275158>
- Teng, M. F. (2023b). Effectiveness of captioned videos for incidental vocabulary learning and retention: the role of working memory. *Computer Assisted Language Learning*, 38(1–2), 206–234. <https://doi.org/10.1080/09588221.2023.2173613>
- Teng, M. F. (2025). Incidental learning of multiword expressions from bilingual captioned videos for language minority students. *The JALT CALL Journal*, 21(1). <https://doi.org/10.29140/jaltcall.v21n1.1201>
- Teng, M. F., & Mizumoto, A. (2023). The role of spoken vocabulary knowledge in language minority students' incidental vocabulary learning from captioned television. *Australian Review of Applied Linguistics*, 46(2), 253–278. <https://doi.org/10.1075/aral.22033.ten>
- Teng, M. F., & Cui, Y. (2025). Second language collocation learning through captioned videos: How do learners' vocabulary knowledge and working memory affect learning? *Computer Assisted Language Learning*, 1–29. <https://doi.org/10.1080/09588221.2025.2497495>
- Valentini, A., Ricketts, J., Pye, R., & Houston-Price, C. (2018). Listening while reading promotes word learning from stories. *Journal of Experimental Child Psychology*, 167, 10–31. <https://doi.org/10.1016/j.jecp.2017.09.022>
- Vanderplank, R. (1988). The value of teletext subtitles in language learning. *English Language Teaching Journal*, 42(4), 272–281. <http://search.proquest.com/docview/58186229/>
- Vanderplank, R. (2016). *Captioned media in foreign language learning and teaching: Subtitles for the deaf and hard-of-hearing as tools for language learning*. Palgrave Macmillan. <https://doi.org/10.1057/978-1-137-50045-8>
- Vanderplank, R. (2019). Video and informal language learning. In M. Dressman & R. W. Sadler (Eds.), *The handbook of informal language learning* (pp. 181–201). Wiley-Blackwell. <https://doi.org/10.1002/9781119472384.ch12>
- Vanderplank, R., & Teng, M. F. (2024). Intentional vocabulary learning through captioned viewing: Comparing Vanderplank's 'cognitive-affective model' with Gesa and Miralpeix. In M. F. Teng, A. Kukulska-Hulme, & J. G. Wu (Eds.), *Theory and practice in vocabulary research in digital environments* (1st ed., pp. 10–24). Routledge. <https://doi.org/10.4324/9781003367543>
- van Zeeland, H., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, 41(3), 609–624. <https://doi.org/10.1016/j.system.2013.07.012>
- Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, 61(1), 219–258. <https://doi.org/10.1111/j.1467-9922.2010.00593.x>
- Vu, D. V., & Peters, E. (2020). Learning vocabulary from reading-only, reading-while-listening, and reading with textual input enhancement: Insights from Vietnamese EFL learners. *RELJ Journal*, 53(1), 85–100. <https://doi.org/10.1177/0033688220911485>

- Vygotsky, L. S. (1978). Interaction between learning and development. In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in society: The development of higher psychological processes* (pp. 79–91). Harvard University Press.
- Vygotsky, L. S. (1998). *The collected works of L. S. Vygotsky: Vol. 5. Child psychology* (R. W. Rieber, Ed.). Plenum Press.
- Wang, A. & Pellicer-Sánchez, A. (2022). Incidental vocabulary learning from bilingual subtitled viewing: An eye-tracking study. *Language Learning*, 72(3), 765–805. <https://doi.org/10.1111/lang.12495>
- Watanabe, Y. (1997). Input, intake, and retention: Effects of increased processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition*, 19(3), 287–307. <https://doi.org/10.1017/S027226319700301X>
- Webb, S., (2020). Incidental vocabulary learning. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 225–239). Routledge.
- Webb, S., & Chang, A. C. S. (2015). How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *Studies in Second Language Acquisition*, 37, 651–675. <https://doi.org/10.1017/S0272263114000606>
- Webb, S., & Rodgers, M. P. H. (2009). Vocabulary demands of television programs. *Language Learning*, 59, 335–366. <https://doi.org/10.1111/j.14679922.2009.00509.x>
- Webb, S. (2015). Extensive viewing: Language learning through watching television. In D. Nunan & J. C. Richards (Eds.), *Language learning beyond the classroom* (pp. 175–184). Routledge. <https://doi.org/10.4324/9781315883472-24>
- Wi, I., & Boers, F. (2024). Sequential use of L1 and L2 captions: Exploring the benefits for vocabulary acquisition. *TESOL Quarterly*, 58(1), 511–521. <https://doi.org/10.1002/tesq.3243>
- Winke, P., Gass, S., & Syodorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology*, 14(1), 65–86. <http://dx.doi.org/10125/44203>
- Wisniewska, N., & Mora, J. C. (2020). Can captioned video benefit second language pronunciation? *Studies in Second Language Acquisition*, 42(3), 599–624. <https://doi.org/10.1017/S0272263120000029>
- Wong, S. W. L., Lin, C. C. Y., Wong, I. S. Y., & Cheung, A. (2020). The differential effects of subtitles on the comprehension of native English connected speech varying in types and word familiarity. *SAGE Open*, 10(2). <https://doi.org/10.1177/2158244020924378>
- Yang, J. C., & Chang, P. (2014). Captions and reduced forms instruction: The impact on EFL students' listening comprehension. *ReCALL*, 26(1), 44–61. <https://doi.org/10.1017/S0958344013000219>
- Zenith. (2019, October). Global online video viewing and AD spending forecast. Marketing Chart. <https://www.marketingcharts.com/digital/video-110520>
- Zou, D. (2017). Vocabulary acquisition through cloze exercises, sentence-writing and composition-writing: Extending the evaluation component of the involvement load hypothesis. *Language Teaching Research*, 21, 54–75. <https://doi.org/10.1177/1362168816652418>
- Zwier, L. J., & Boers, F. (2023). *English L2 vocabulary learning and teaching: Concepts, principles, and pedagogy*. Routledge.

Author Biographies

Injung Wi is a doctoral candidate in Applied Linguistics at Western University, Canada, where she also earned an MPed in TESOL and an MA in Applied Linguistics. She has nine years of experience teaching English to EFL learners. Her research interests include second language vocabulary acquisition through audiovisual input and the use of on-screen text.

Frank Boers is Professor in Applied Linguistics and TESOL at Western University, Ontario. His early research was in the area of lexical semantics, but his interests shifted after he became a language teacher and then a teacher trainer. Since the 2000s, he has published mostly on matters of second language education.