Language Teaching (2022), **55**, 163–192
doi:10.1017/S0261444821000501

CAMBRIDGE UNIVERSITY PRESS

STATE-OF-THE-ART ARTICLE

Second or foreign language learning through watching audio-visual input and the role of on-screen text

Maribel Montero Perez

Ghent University, Gent, Belgium
Email: Maribel.monteroperez@ugent.be

(Received 3 December 2021; accepted 4 December 2021)

Abstract

This article discusses research into the role of audio-visual input for second language (L2) or foreign language learning. It also addresses questions related to the effectiveness of audio-visual input with different types of on-screen text such as subtitles (i.e., in learners' first language) and captions (i.e., subtitles in the same language as the L2 audio) for L2 learning. The review discusses the following themes: (a) the characteristics of audio-visual input such as the multimodal nature of the input and vocabulary demands of video; (b) L2 learners' comprehension of audio-visual input and the role of different types of on-screen text; (c) the effectiveness of audio-visual input and on-screen text for aspects of L2 learning including vocabulary, grammar, and listening; and (d) research into L2 learners' use and perceptions of audio-visual input and on-screen text. The review ends with a consideration of implications for teaching practice and a conclusion that discusses the generalizability of current research in relation to suggestions for future research.

1. Introduction

Research into the role of video for second language (L2) learning dates back to the 1980s (e.g., Price, 1983; Vanderplank, 1988). While some of the questions from these early studies (i.e., Do L2 learners acquire vocabulary while watching video?) still inspire current work, the way in which we access and use audio-visual input has evolved enormously in the past decade. For instance, media consumption is no longer restricted to TV and DVD but also takes place through video-based platforms such as YouTube and Vimeo, social media, and the increasingly popular video streaming platforms (e.g., Netflix, Hulu, HBO), which can be accessed from a variety of devices such as laptops, tablets, and mobile phones. Video-based websites as well as social media have also contributed to the emergence of new digital text genres such as TED-talks, video tutorials, so-called react videos (i.e., videos in which there is a split-screen with a source-video and the reaction of a person while watching that video), or interactive videos (e.g., on Netflix) in which viewers make choices for characters while watching a movie or series. In addition, options to modify L2 audio-visual input such as on-screen text in the native language (i.e., subtitles) or in the target language (i.e., captions) have become easier to access thanks to, for instance, YouTube, and streaming services.[1]

This article provides a review of research into the role of video as input for L2 learning. The starting point for this article is Robert Vanderplank's state-of-the-art article, 'Déjà vu? A decade of research on language laboratories, television and video in language learning', published in this journal in 2010. In the article, Vanderplank discussed research on the role of television and video in L2 learning as well as several studies on the value of video with captions and subtitles for promoting L2 comprehension and learning. One decade later, there is an exponential increase in studies on the effects of video on L2 learning. There are now numerous articles reporting on primary research into the role of video and TV for L2 learning in journals on second language acquisition (SLA) and computer assisted language

© The Author(s), 2022. Published by Cambridge University Press

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

learning (CALL). Furthermore, the topic has led to the publication of several books and edited volumes on the role of film for L2 learning (Herrero & Vanderschelden, 2019), the use of video in language learning activities (e.g., Goldstein & Driver, 2015) and L2 teaching (e.g., Sherman, 2010), and the role of subtitled video (Gambier et al., 2015; Ghia, 2012) and captioned video (Teng, 2021; Vanderplank, 2016a) in L2 learning. There is also a meta-analysis on captioned video (Montero Perez et al., 2013) as well as several comprehensive reviews on (captioned) video for L2 learning (e.g., Caruana, 2021; Vanderplank, 2016b, 2020), captioned video for L2 listening (Yeldham, 2018) or on the role of subtitled video for L2 learning (Black, 2021). In addition, the role of audio-visual input has inspired conferences (e.g. Subtitles and language learning, Pavia, 2012), thematic colloquia at international conferences (e.g., Multimodal input and L2 learning, Eurosla 2018; The potential of audio-visual input for language learning, *Second Language Research Forum*, 2018), and two recently published special issues: Video and language learning (Montero Perez & Rodgers, 2019 for *The Language Learning Journal)* and Language learning from multimodal input (Peters & Muñoz, 2020 for *Studies in Second Language Acquisition).* Together, these publications and initiatives make a synthesis article timely and important.

To identify a relevant body of publications for this review on audio-visual input for L2 learning, five criteria were used. The first criterium concerns the focus of this article: studies that have used auditory input in the L2[2] with dynamic visuals such as television, film, and video as input for L2 learning. In this article, audio-visual input and video will be used interchangeably in order to refer to the use of video as input for L2 learning. Other uses of video in SLA and CALL research such as videoconferencing (e.g., Satar, 2016), video-based gaming (e.g., deHaan et al., 2010), and video-based Computer-Mediated Communication (Kato et al., 2016), or contexts in which L2 learners create videos, add voiceover or subtitles to videos (e.g., Talaván & Avila-Cabrera, 2015), and educational video use in online learning environments such as MOOCs (e.g., Laaser & Toloza, 2017) will not be included. It does not consider studies in which audio-visual input is defined in terms of a combination of audio and static imagery. Second, with regard to the timeframe that was used in the selection process, this review starts where Vanderplank's article stopped, that is, in 2009–2010. However, where necessary, links to earlier work will be established. For instance, this review includes numerous studies into the effects of audio-visual input with on-screen text that build on pivotal work carried out in the 1980s and 1990s (e.g., Borrás & Lafayette, 1994; Neuman & Koskinen, 1992). Third, this review includes studies on the role of audio-visual input for learners with different proficiency levels and for learners of different age groups. However, studies in which learners did not yet develop basic reading skills were not included (e.g., Wong & Samudra, 2019 on acquisition of vocabulary from watching L2 educational media in preschool children). Fourth, even though most studies have focused on English as L2, this review also includes findings for other target languages such as, for instance, French and Russian. Fifth, journal articles, monographs, and book chapters written in English were included. Ph.D. theses, conference presentations as well as publications not written in English were not considered.

The main goal of the current review is to provide a synthesis of recurring themes and issues in the literature on video for L2 learning and the role of on-screen text. The article is structured as follows: Section 2 focuses on key characteristics of audio-visual input. Next, Section 3 presents studies that focus on L2 learners' video comprehension and the role of different types of on-screen text. The following sections deal with the effects of audio-visual input and different types of on-screen text on L2 learning including vocabulary (Section 4), grammar and listening (Section 5). Section 6 addresses questions related to learners' perceptions and use of audio-visual input for L2 learning, whereas Section 7 provides pedagogical implications.

2. Characteristics of audio-visual input

2.1 The multimodal nature of audio-visual input

Video offers a combination of input modes including auditory (i.e., spoken input) and visual input. Visual input consists of dynamic imagery and can also include written on-screen text. Table 1 provides

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

Table 1. Types of on-screen text in studies published in SLA and CALL journals

Language in audio	Language in on-screen text	Length and information density in on-screen text	Terminology
L2	L2	Verbatim transcription	Intralingual subtitles; L2 subtitles; (full) **captions**
L2	L1	Translation of L2 audio (see Szarkowska et al., 2021 for a description of subtitle quality). Subtitles may imply information loss or impact information density owing to formatting conventions for subtitling (e.g., 40 characters per line)	Interlingual subtitles, L1 subtitles, **subtitles**
L2	L2	Only keywords are presented in the on-screen text area	Keyword captions
L2	L2	Parts of sentences are presented in the on-screen text area	Partial captions
L2	L2	Verbatim transcription but specific words or structures are made salient through, for instance, bolding, underlining, and highlighting	Captions with textual or typographical enhancement
L2	L1 and L2	L1 and L2 are presented simultaneously in on-screen text area (e.g., L1 in upper line, L2 in lower line)	Double subtitles, dual subtitles, **bilingual subtitles**
L2	L1 and L2	Verbatim transcription in L2. Difficult words or multiword units are glossed and provide access to the corresponding L1 translation	Glossed captions
L1	L2	On-screen text in L2 translation	Reversed subtitles

Note. When multiple terms are available, the term that will be used in this article appears in bold.
Reversed subtitled input is not considered in this review because it does not expose learners to spoken L2 input.

an overview of different types of on-screen text that have been studied in previous research. Because audio-visual input combines different input modes, it has also been characterized as a multimodal type of input. Multimodality has been defined as:

> . . . the use of several semiotic modes in the design of a semiotic product or event, together with the particular way in which these modes are combined – they may for instance reinforce each other ('say the same thing in different ways'), fulfill complementary roles [e.g. when pictures complement a story in a children's book] or be hierarchically ordered [e.g. dramatic music in an action movie]. (Kress & Van Leeuwen, 2001, p. 20).

Drawing on the importance of multimodal interaction and communication in our daily life, The Douglas Fir Group (2016) has defined L2 learning as multimodal and semiotic because language users and learners rely on a wide array of semiotic resources to construct meaning. In an everyday conversation, for instance, we use contextual information, visuals (e.g., face expressions), intonation, and so forth, to extract and build meaning. It has also been argued that multimodality has reached a 'qualitatively new level' (The Douglas Fir Group, 2016, p. 22) thanks to technological advances that have promoted the use of digital tools in daily life (e.g., online courses, videoconferencing).

The increasingly important role of multimodality may also impact how we approach texts. Serafini (2012) proposed a reconceptualization of the reader as a 'reader-viewer' who uses specific skills to deal with multimodal texts. For instance, the reader-viewer is an interpreter who needs to construct

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

meaning from the different aspects included in a multimodal text. According to Paivio (1986, p. 53), human cognition:

> . . . has become specialized for dealing simultaneously with language and with nonverbal objects and events. Moreover, the language system is peculiar in that it deals directly with linguistic input and output (in the form of speech or writing) while at the same time serving a symbolic function with respect to nonverbal objects, events, and behaviours.

Drawing on Paivio's dual coding-theory, Mayer (2001) proposed the cognitive theory of multimedia learning. This theory posits that there are two separate channels for information processing that refer to presentation modes of information (i.e., verbal or nonverbal) or to sensory modalities (i.e., visual or auditory). According to the cognitive theory of multimedia learning, the simultaneous activation of these channels is expected to foster learning as indicated by the multimedia learning principle that holds that 'people learn better from words and pictures than from words alone' (Mayer, 2001, p. 63). Even though the cognitive theory of multimedia learning was initially developed and studied in the context of first language (L1) science learning, it has also been linked explicitly to L2 learning contexts (Mayer et al., 2014). In addition, several of its principles, such as the spatial and temporal contiguity principle (see Section 3.1) and the redundancy principle (see Section 3.2), have been used in studies on L2 learning through watching audio-visual input.

2.2 Authenticity of language use in audio-visual input versus naturally occurring input

It has been asked whether language use in TV-programs and movies reflects real or authentic language use as compared with naturally occurring language use. This question has inspired corpus-driven research that is based on analyses of language use in specific corpora such as, for instance, The Sydney Corpus of Television Dialogue (Bednarek, 2018), The TV Corpus, and The Movie Corpus (Davies, 2021).

Overall, findings seem to indicate that television dialogue presents linguistic features of natural conversation even though it is often based on scripted language (e.g., language use in the TV sitcom *Friends*). For instance, it has been found that TV series are comparable to natural language use in terms of lexicogrammatical features (Bednarek, 2018) and provide access to pragmatic information such as data on compliments and compliments responses (Rose, 2001). TV dialogue also reflects colloquial and informal language use (i.e., spontaneous talk) and contains new expressions that reflect contemporary language use (Bednarek, 2018).

The degree to which scripted language use presents natural language use may, however, depend on genre. Using Bibers' Multidimensional analysis, Al-Surmi (2012) found that a sitcom series is closer to natural conversation than soap operas for some of the dimensions analyzed (e.g., expression of argumentation or persuasion), whereas soap operas were found to be closer to naturally occurring language use for narrative discourse (e.g., narrating events). However, sitcoms can also diverge from natural conversation in multiple ways. For example, language use in a corpus based on transcripts from the sitcom *Friends* contained fewer interruptions or overlaps (Quaglio, 2010) than real language use. TV-series may also contain overrepresentations of certain features such as routine formulae or items such as 'tell me what' (Bednarek, 2018). Further, it has also been argued that TV dialogue contains highly informal and non-standard elements of language use. In this respect, Werner (2020) argued that TV dialogue represents an interesting approach to illustrate grammaticality and informal spoken usage across varieties, which may contribute to learners' language awareness. However, research in this particular area is scarce and does not yet reflect the diversity in types of audio-visual input that not only include TV programs but also YouTube videos or video-based social media (see Terantino, 2011, for a review on YouTube for foreign language learning).

Studies have also examined the authenticity of vocabulary use in TV series. For instance, Lin (2014) compared the distribution of formulaic sequences in a corpus of internet television to the spoken

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

British National Corpus (BNC). She reported that the use of formulaic sequences in an internet television corpus reflects the frequency of formulaic sequences in everyday use. This highlights the potential of TV for learning formulaic language, as has also been found in experimental research (see Section 4). Sockett (2012) investigated which four-word combinations occurred most frequently in a corpus of five television series. It was found that the 50 most frequently occurring four-word clusters correlated with real-life spoken English (as based on BNC data). Scheffler et al. (2021) investigated vocabulary use in a corpus of an animation series (*Peppa Pig*) for pre-primary children. Results showed that 80% of the tokens in *Peppa Pig* stem from the 2,000 most frequent words in the BNC and that the 25 most frequently used words in the animation series are comparable to the 25 most frequent words in the spoken component of the BNC (i.e., high frequency pronouns, articles and prepositions). Together, these figures corroborate Vanderplank's definition of TV as 'a very verbal medium' (Vanderplank, 1993) by indicating that TV-series and programs reflect real language use and that exposure to these types of materials may hold great potential for L2 leaning.

2.3 Vocabulary demands of audio-visual input

Several corpus-based studies have sought to determine the vocabulary demands of different types of audio-visual input to identify the vocabulary size needed for adequate comprehension. Webb and Rodgers (2009a, 2009b) analyzed the transcripts of, respectively, 318 movies and 88 TV programs such as news and TV series (e.g., sitcom, drama, sci-fi) in order to determine the vocabulary size needed for a 95% and 98% lexical coverage, that is, the amount of words in the input that a learner needs to know. These percentages had been defined as threshold percentages for adequate global and detailed comprehension in the context of written input (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010).

Generally speaking, Webb and Rodgers found that knowledge of 3,000 word families (plus proper nouns and marginal words) is needed for a 95% coverage of movies (Webb & Rodgers, 2009a) and TV programs (Webb & Rodgers, 2009b). Yet, depending on the genre, figures of 2,000 to 4,000 word families (plus proper nouns and marginal words) have been put forward for 95% coverage. Studies on written input had found that knowledge of 4,000–5,000 most frequent word families is needed to achieve 95% coverage of written input (Laufer & Ravenhorst-Kalovski, 2010). A comparison of vocabulary demands of written and audio-visual input thus seems to indicate that vocabulary demands of audio-visual input are lower than for reading. In addition, there are different conditions that may lower the vocabulary load of video further such as, for instance, a narrow viewing approach.

Narrow viewing, which consists of watching related episodes of a TV-show (e.g., an entire season of a TV-series) or videos that are related to the same topic, can lower the vocabulary load for L2 learners. Results from a corpus-driven study (Rodgers & Webb, 2011) that compared the transcripts of 288 television episodes revealed that related TV programs tend to contain fewer word families than unrelated TV-programs. Rodgers and Webb also found that lower-frequency word families tend to reoccur more often in related than in unrelated programs. Narrow viewing therefore has the potential to reduce the vocabulary load, which may subsequently contribute to L2 learners' video comprehension. In addition, related programs may stimulate learners' use of contextual and prior knowledge during the comprehension process because they become increasingly familiar with characters, plots, and settings.

Corpus-driven studies have also suggested that previewing activities can lower the vocabulary demands of audio-visual input. Webb (2010) found that pre-learning the ten most frequently occurring topic-related low-frequency word families can increase learners' lexical coverage of a television programme from 0.70% to 3.91% and thus has the potential to improve learners' comprehension of television programmes.

Finally, it has been argued that learners might achieve adequate comprehension of the input with coverage percentages lower than 95% when images are supportive of the meaning (Webb & Rodgers, 2009a). Durbahn et al. (2020) measured L2 learners' comprehension of an English documentary with a test that consisted of questions that were based on the audio, on the images, and on audio + images. They found that comprehension increases as learners' lexical coverage increases. In addition, when

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

learners can use images to answer the question, they rely less on their vocabulary knowledge than when they have to rely on audio alone. It should also be noted that their data did not reveal a threshold percentage for adequate comprehension. In sum, their findings suggest that imagery can be used to enhance comprehension and that lower levels of lexical coverage are needed in audio-visual input in which there is an important audio-imagery correlation.

However, not all videos and genres contain highly supportive imagery. Cross (2011) analyzed the audio-visual correspondence in news videotexts using the four-category coding system of Walma van der Molen (2001, as referenced in Cross, 2011). Visuals were coded as talking head (i.e., no imagery beyond the person talking), direct (i.e., audio and visuals provide the same meaning), indirect (i.e., audio and visuals are not directly related), and divergent (i.e., audio and visuals are not related or there is a mismatch). His findings indicated that visuals in news videotexts are mainly talking heads, followed by indirect or divergent visuals. Images that were directly supportive of the audio were rare in the videos analyzed. Rodgers (2018) explored the potential of imagery for clarifying the meaning of unknown vocabulary in narrative TV genres (a fantasy/adventure TV series) and a documentary. It was found that both genres contained images that were supportive of the meaning of difficult words and that images in documentaries were more related to target vocabulary than narrative TV genres. The author links these findings to the spatial and temporal contiguity principle of the multimedia learning theory, which posits that 'students learn better when corresponding words and pictures are presented near rather than far from each other' (Mayer, 2001, p. 81) and that words and corresponding pictures should be 'presented simultaneously rather than successively' (Mayer, 2001, p. 96). In light of these principles, the images in documentaries were also found to co-occur (i.e., within a time frame of 5 seconds before or after) with the aural word form in 72% of the selected target words, which may promote information processing and learning (see Section 4.5). Together, these findings suggest that supportive imagery can lower the vocabulary demands of audio-visual input and that they may help learners in their meaning-making process, which makes the audio-image link an important criterium to consider when selecting audio-visual materials for L2 learning.

3. Comprehension of L2 audio-visual input and the role of on-screen text

3.1 L2 learners' video comprehension and the effects of captions

Already in 1988, Vanderplank argued that on-screen text such as captions (see Table 1) has the potential to 'unlock' input that would otherwise be too challenging (Vanderplank, 1988, p. 278). This was also found by Price (1983) who reported on a large-scale experiment comparing the effects of watching video with and without captions on 500 university students' content comprehension. Results revealed that captions yielded superior comprehension scores than video without captions and this was independent of language and educational background, social, and linguistic variables. While there is no detailed research report, it has inspired numerous researchers to investigate L2 learners' comprehension of audio-visual input as well as ways to promote their comprehension of audio-visual input by means of captions.

So far, one meta-analysis (Montero Perez et al., 2013), which was based on data of 15 studies published between 1989 and 2012, has investigated the effectiveness of captioned video as compared with uncaptioned video for L2 comprehension. A significantly large effect of captioning ($g = 0.99$) was found, which indicates that participants who were exposed to captions scored significantly better on comprehension tests than participants who were exposed to video without captions. More recent results are consistent with these findings (e.g., Majuddin et al., 2021). The meta-analysis also revealed that the positive effects of captions were found irrespective of learners' proficiency level.[3] However, the effect of captions was moderated by test type and the largest effects were found for captions when a receptive comprehension test was used (e.g., multiple choice) as compared with a productive test (e.g., recall test). The facilitative effect of captions on L2 comprehension has been explained by the fact that captions seem to assist L2 learners' 'attentional control' (Gass et al., 2019, p. 98) during input processing because they help learners to map the written and auditory trace. This may subsequently assist

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

learners' segmentation and decoding of the speech stream (see Section 5.2), which provides them with more time to process and interpret the meaning of the incoming information.

Using eye tracking, Kruger and Steyn (2014) explored the relationship between L2 learners' comprehension of video and their captioning use. Participants watched six English educational videos in a captioning or no captioning group over a two-week period and took a comprehension test two weeks after the treatment. While no differences between the captioning and no captioning groups were found on the comprehension test, caption-reading behavior (i.e., learners' fixations on the captions) was positively and significantly linked to performance on the comprehension test. The authors therefore argue that learners' individual reading behavior should be considered in order to understand the global effect of captions on comprehension.

While numerous studies have revealed a positive effect of captions as compared with video without captions on L2 comprehension, the redundancy principle of the cognitive theory of multimedia learning would state that presenting the same information in different forms (e.g., audio and corresponding on-screen text) negatively affects learning. In addition, using on-screen text with video might lead to extraneous processing (Sweller, 2005) in the visual channel that needs to process both the captions and the imagery, which could lead to a split-attention effect in the visual/pictorial channel. While these effects have been documented in numerous empirical studies dealing with science learning in learners' L1 (see Adesope & Nesbit, 2012, for a meta-analysis on the redundancy effect), the aforementioned findings with L2 learners indicate that redundancy operates differently in L2 learning contexts. Mayer et al. (2020) therefore propose 'the subtitle principle' that indicates that redundant information in the form of on-screen text may be helpful for comprehension in L2 learning contexts where learners have word identification problems.

3.2 Comprehension of (captioned) video under different viewing conditions

There are several viewing conditions that might influence L2 learners' comprehension and in some cases the effectiveness of captioning. However, it should be noted that results remain provisional because of the limited number of studies in this area.

The number of viewings may influence L2 learners' comprehension. Majuddin et al. (2021) found that viewing a video twice led to significantly higher comprehension scores than watching once, but this did not seem to interact with captioning effectiveness. Winke et al. (2010) argued that when using repeated viewings, the order of viewing with captions (captions during the first viewing or during the second viewing) could play a role. They suggest that captions can be considered during the first viewing of the clip when the orthography of a learner's mother tongue is close to the script of the L2 (e.g., Spanish-speaking learners of English). Their data provide indications that captions might be more effective in a second viewing when the L1 and L2 represent greater orthographic distance (e.g., English-speaking learners of Arabic). However, it should be noted that the differences between conditions did not reach significance and that more research is needed to uncover whether learners' L1 modulates the effectiveness of captioned video.

Mayer et al. (2014) investigated the role of captioning in relation to the pace of the video. They found that captioning did not lead to superior comprehension (compared with no captions) of a fast-paced narration. Fast-paced was defined as a video with 142 words per minute that is, however, comparable to naturally occurring speech containing approximately 150 words per minute. Drawing on these findings, they argue that the pace of the video should be slow enough in order to give learners time to process the on-screen text and avoid working memory overload (see, for instance, Frumuselu, 2018 and Liao et al., 2020 on cognitive load and subtitles).

Narrow viewing, which consists in watching several related episodes (see section 2.3) may have an effect on captioning effectiveness. Using a multiple-session design, Rodgers and Webb (2017) investigated the effects of watching ten episodes of a TV-series (*Chuck*) in a captioned and uncaptioned condition on intermediate L2 learners' comprehension. Participants watched ten 42-minute episodes (one per week) of the series and completed a comprehension test after each episode. It was found that all participants, irrespective of condition, significantly improved their comprehension scores from the

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

first to the tenth episode, which indicates that regularly watching audio-visual input is beneficial for L2 learners' listening comprehension. Findings also showed that learners with captions had significantly higher scores on the comprehension test than learners in the uncaptioned video for the first episode and for the episodes that were the most difficult ones. This seems to indicate that the large effect of captions in single-session treatments with short video exposure may be different in contexts with more sustained exposure to captioned video (i.e., longer clips over a longer period) and that input that is serial in nature may also facilitate comprehension irrespective of captioning availability because of learners' growing knowledge of the series (i.e., characters, contexts, storylines).

3.3 Video comprehension and the length of captions

Several studies have investigated the question of whether captions should be full-length or only represent words that are important for the meaning of the excerpt, hence substantially reducing L2 learners' reading effort while watching video. These studies draw on the work of Guillory (1998) who found that beginning-level learners achieved similar comprehension scores in a full or keyword captioning group (representing only 14% of the full captions). She therefore argued that keywords are an interesting alternative to captions as they require less reading and have the potential to draw learners' attention to specific ideas in the video.

Building on these findings, Montero Perez et al. (2014b) further explored the potential value of keyword captions. They found, however, that intermediate university-level learners who watched short videos with full captions outperformed the keyword group (who saw 14% of the words in the captions) and no captioning group on the global comprehension questions. Similar findings regarding the effects of keyword captions were reported by Teng (2019a) in a study with primary school learners. Additionally, Lee et al. (2021) took into account learners' self-reported caption reliance (i.e., how much they rely on captions while watching video) when comparing the effects of full, partial, and no captions on L2 learners' comprehension. It was found that learners in a high captioning reliance group (see Section 6.2) performed best on a comprehension test after watching an 11-minute TED-talk with full captions (as compared with partial and no captions). However, learners in the low captioning reliance group were found to score best under a partial captioning condition. This raises questions about the role that individual differences and preferences might play in captioning effectiveness.

So far, results on the role of partial or keyword captions are inconclusive. Moreover, several potential caveats related to the use of keyword captions can be identified. First, the selection of words to show in the keyword caption mostly relies on manual selections. Yet, these selections, as well as the creation of keyword captions, are time-consuming – which may hinder the actual pedagogical use of such a tool. However, by using speech technology, Mirzaei et al. (2017) were able to develop a partial (here, 30% of the caption text) and synchronized captioning tool. The words that appeared in the captions were automatically selected as a function of three parameters: speech rate, word frequency, and word specificity (i.e., technical vocabulary). It was found that participants in the partial and full captioning groups outperformed the control group, and that learners who had partial captions did not differ significantly from the full captioning group on a comprehension test, irrespective of their proficiency level. A second caveat is related to learners' perceived usefulness of keyword captions. It has been found that learners perceived keywords as distracting and that they hampered their listening experience (Montero Perez et al., 2014). Finally, the percentage of words covered seems to vary across studies (e.g., from 14 to 30%) and may play an important role in explaining the mixed findings.

3.4 Video comprehension and the language(s) of on-screen text: Captions, subtitles, and bilingual subtitles

Whereas captions provide access to the written account of the L2 input, subtitles provide a translation of the audio in learners' L1. Subtitles may therefore assist L2 learners' comprehension of audio-visual input irrespective of L2 proficiency. The differential effects of subtitles and captions have been

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

investigated in short clips (for early work on this topic, see Markham et al., 2001) as well as full-length episodes of a TV series (e.g., Birulés-Muntané & Soto-Faraco, 2016; Vulchanova et al., 2015). Findings point to an overall positive effect of subtitles on comprehension with learners who have access to subtitles outperforming learners in video only conditions (Vulchanova et al., 2015) as well as in captioned and video only conditions (e.g., Bianchi & Ciabattoni, 2008; Birulés-Muntané & Soto-Faraco, 2016) even though findings did not always reach significance levels (e.g. Liao et al., 2020). Pujadas and Muñoz (2020) found that adolescent English as a Foreign Language (EFL) learners in a subtitling group also outperformed participants in a captioning group in an extensive viewing treatment in which they watched 24 episodes of a TV-series. In a study comparing different proficiency groups (i.e., beginning, intermediate, advanced learners), Lavaur and Bairstow (2011) found that beginning adolescent L2 learners benefitted most from subtitles as compared with captions and video only. No differences between the types of on-screen text were found for intermediate learners, whereas advanced learners benefitted most from video without on-screen text. Together, these findings suggest that subtitles can be considered effective for promoting L2 learners' comprehension of audio-visual input and they seem to be especially relevant for making L2 audio-visual input available to beginning L2 learners who do not yet have sufficient reading speed in the L2 to deal with captions (Vanderplank, 2016a) or who do not yet have enough L2 knowledge to process captions efficiently (Muñoz, 2017).

The role of bilingual subtitles (see Table 1 for a definition) has only recently attracted researchers' attention. While this type of subtitling is less commonly found, it is often used in countries with multiple official languages (e.g., French and Dutch subtitles in movie theatres in Belgium) and they are increasingly being used in China (Liao et al., 2020). The question whether bilingual subtitles are superior to other types of on-screen text for improving comprehension remains largely unanswered. Wang (2019), for instance, reported that short clips from an English TV series with bilingual subtitles were more effective than having the clips with captions for one class of freshman students. However, no differences were found for a different group of freshman students even though all participants were (higher) intermediate learners of English. Lwo and Chia-Tzu Lin (2012) as well as Liao et al. (2020) did not find differences between a captioning, subtitling, bilingual subtitling, and video only group. One potential explanation for the lack of differences between these different conditions can be found in the eye-movement data reported in Liao et al. (2020). More specifically, they found that bilingual captions did not seem to significantly alter L2 learners' viewing behavior, and that they tended to pick one language of the bilingual subtitle as main source of information. Participants thus spent similar amounts of time on the L1 text in the bilingual condition as on the text in the subtitle only group. Considering limitations in working memory capacity and the cognitive load theory, questions may be asked as to whether and when the simultaneous presentation of subtitles and captions is needed for L2 learners.

4. Learning new vocabulary from watching L2 audio-visual input and the role of on-screen text

4.1 Vocabulary learning from video

During the last decade, there has been a surge of studies on vocabulary learning through watching L2 audio-visual input. Since this type of vocabulary learning takes place through contextual exposure, it can also be labelled contextual word learning (see Elgort et al., 2018 on contextual word learning from written input). Lin and Siyanova-Chanturia (2015) argued that internet television may be an ideal approach to contextually learn new vocabulary because it may provide multimodal contextual cues. More particularly, this means that L2 learners can use verbal (spoken and/or written) as well as nonverbal (i.e., imagery) information to infer the meaning of new words.

To measure vocabulary learning that takes place through contextual exposure, previous studies have mainly relied on tests that tap into form-meaning mappings (see Table 2 for an overview of frequently used tests).

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

Table 2. Frequently used types of vocabulary tests in research on L2 vocabulary learning from video (single words and multiword units)

	Form	Meaning
Receptive	*Form recognition* e.g., Do you know this word? / Have you seen this word before? Yes/No (e.g., Feng & Webb, 2020)	*Meaning recognition* e.g., Select the correct L1 translation from 4 options (e.g., Peters & Webb, 2018)
Productive	*Form recall* e.g., Provide the L2 form for L1 meaning (e.g., Puimège & Peters, 2020)	*Meaning recall* e.g., Provide a synonym, description or L1 translation for L2 form (e.g., Sydorenko, 2010)

Previous research[4] has shown that video can be considered a source of vocabulary learning. Results indicate that watching a single TV program or a full-length episode has the potential to stimulate L2 learners' incidental acquisition[5] of single words. In Peters and Webb (2018), it was found that watching an English 1-hour documentary had a significant effect on L2 learners' incidental vocabulary learning as measured by means of a meaning recognition and meaning recall test with approximately four words learned from pretest to posttest. Montero Perez (2020) investigated a similar question but manipulated a French video to include 15 pseudowords, that is, the targets for incidental learning. Learning occurred at the level of form and meaning recognition with approximately six words learned on the form and three words learned on the meaning recognition tests. Other studies have expanded on these findings by showing that learners do not only acquire single words but also formulaic sequences from watching audio-visual input. Significant incidental learning of the form (Puimège & Peters, 2019a, 2019b) and meaning (Puimège & Peters, 2020) of formulaic sequences was found after watching an English documentary. Learning gains were influenced by several word-related factors including corpus frequency, collocate-node relationship (i.e., the combination of word classes or parts of speech within the formulaic sequence), semantic transparency (i.e., the degree to which the meaning of the formulaic sequence can be inferred from the constituent parts), as well as the association strength of the constituent components of the formulaic sequence.

Another question that has been asked is whether audio-visual input is comparable to written input in terms of the amount of contextual vocabulary learning that can take place (Webb, 2015). Arndt and Woore (2018) compared incidental vocabulary learning of six pseudowords through YouTube videos and written blog posts. Vocabulary learning was measured by means of multiple tests targeting recall of orthographic form, meaning, and grammatical function as well as recognition of grammatical function and meaning. It was found that videos and blog posts were comparable in terms of learning gains but blog posts led to significantly higher gains on orthographic knowledge presumably because learners saw the written form of the pseudowords in the blog. Similarly, Feng and Webb (2020) compared incidental vocabulary learning through three modalities: written input, auditory input, and audio-visual input and found that all modalities led to vocabulary learning gains on a meaning recognition test. However, no differences were found between the groups, which seems to indicate that exposure to different input modes can lead to comparable incidental learning gains. As indicated by the authors, it is somewhat surprising that audio-visual input did not result in larger learning gains considering the potential value of imagery for word learning. They argue that this might be because their participants (i.e., EFL learners in China) were less familiar with the use of audio-visual input as learning materials.

4.2 Vocabulary learning from captioned video

Many studies have focused on whether vocabulary learning from video can be enhanced by means of captions. In a meta-analysis on captioned versus uncaptioned video (Montero Perez et al., 2013), which was based on data from ten studies published before 2013, an overall large effect of captions

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

on vocabulary acquisition ($g = 0.87$) was found, irrespective of the type of vocabulary test used and the proficiency level of the learners. This finding indicates that watching captioned video results in significantly larger vocabulary gains than uncaptioned video, which has been attributed to the fact that captions help learners to segment and decode the auditory input (see section 5.2). This may, in turn, facilitate L2 learners' noticing of unknown words in the input. The superiority of captioned video as compared with uncaptioned video was further echoed in more recent findings for learning different aspects of word knowledge of single words such as form recognition (e.g., Chen et al., 2018; Peters, 2019; Sydorenko, 2010), meaning recognition (e.g., Teng, 2019b), and meaning recall (e.g., Mohd Jelani & Boers, 2018; Teng, 2019b), as well as primary school learners' acquisition of verb-noun collocations (Teng, 2019c).

While there is robust evidence to support the effectiveness of captions as compared with video only for L2 vocabulary learning, it has also been argued that the congruency between input and test modality might influence results of captioning effectiveness. More particularly, this means that the use of written vocabulary posttests in captioned video research may have favored participants who had access to captions. In order to address this issue, Sydorenko (2010) presented half of the words in the recognition posttests in aural form and half in written form in a study with beginning learners of Russian. However, no input-test modality congruency effect was found: learners who had access to captions outperformed those in the uncaptioned group on the written and aural prompts of the meaning recognition test. Mohd Jelani and Boers (2018), however, found that learners in the captioned group outperformed learners in the uncaptioned group on a meaning test for the written prompts but not for the aural ones. This seems to suggest that test modality should be taken into consideration in future research on this topic to obtain reliable measurements. Other studies have dealt with this risk by presenting written prompts and their corresponding spoken form simultaneously in vocabulary posttests (e.g., Peters, 2019).

4.3 Vocabulary learning and the role of captions with textual enhancement, keyword captions, and glossed captions

While captions can stimulate learners to link spoken and written word forms while processing the audio-visual input, they do not overtly direct learners' attention to new lexical units. In addition, while contextual clues and imagery can promote L2 learners' acquisition of form-meaning mappings, they do not always provide explicit word meaning information. It has therefore been investigated how captions can be modified to promote noticing and form-meaning mappings.

Different techniques can be considered to foster learners' noticing of vocabulary while watching captioned video. Studies have investigated the effects of making target words or structures in the captions visually salient through highlighting. Captions with highlighted target words were found to be more beneficial than captions without highlighted words for immediate word form recognition and form recall (Cintrón Valentín et al., 2019) as well as meaning recognition (Montero Perez et al., 2014b) and translation (Cintrón-Valentín & Garciá-Amaya, 2021) of single words. However, Cintrón-Valentín and García-Amaya (2021) found that the advantage of highlighting was not maintained in the delayed posttests that were administered two weeks later. As argued by the authors, this might be explained by the fact that learners did not encounter the words in the two-week period preceding the delayed test while opportunities for retrieval practice are crucial to lead to retention.

Majuddin et al. (2021) further expand on the objectives of Cintrón-Valentín and Garciá-Amaya (2021) by investigating the role of typographic enhancement realized through bolding and underlining multiword expressions in captioned video. They found that learners who had captions or captions with typographic enhancement picked up significantly more multiword expressions than a video only group but typographic enhancement did not lead to more gains than captioning without typographic enhancement. One feature that might explain the similar gains in the captioning groups is the length of the target structures used in the study. Since these could consist of up to five words, the effects of typographic enhancement may have been minimized.

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

Studies have also investigated the role of visual salience operationalized through the presence of keyword captions. However, the impact of keyword captions on vocabulary learning remains unclear. While keywords were found to be as effective as captions and captions with highlighted keywords for adult intermediate learners' vocabulary learning (Montero Perez et al., 2014a), captions seemed to be more effective for the learning of collocations than keyword captions in primary school EFL learners (Teng, 2019c).

In order to promote learners' form-meaning mappings, glossed captions can be considered. Glosses in captions can be accessed by clicking on a word in the caption in order to see its meaning such as the L1 translation (see Boers, 2021 for a review on glosses for vocabulary learning). While research on the use of glossing in audio-visual input is scarce, it has been found that glosses that provide access to an L1 translation can result in significantly more form-meaning mappings than full or keyword captioning (Hsieh, 2020; Montero Perez et al., 2018; Teng, 2020). The advantage of glossed captions is that learners have immediate access to a translation and do not need to rely on the availability of contextual cues or imagery to infer word meaning from video. In addition, positive relationships have been found between gloss consultation and vocabulary posttest scores after watching short clips (Montero Perez et al., 2018) and multiple full-length episodes of a TV-series (Fievez et al., 2021), which further strengthen the potential of glossing form the installation of form-meaning mappings.

In sum, different techniques to make vocabulary more salient in the captions (i.e., highlighting, bolding, keywords) have been investigated but they have yielded mixed results. Results regarding the positive effect of glossed captions seem to be more clear-cut despite the limited number of studies in this area.

4.4 Vocabulary learning from subtitled video

Early work on the role of subtitled video for vocabulary learning has been conducted with young L2 learners (e.g., Koolstra & Beentjes, 1999; d'Ydewalle & Van De Poel, 1999) and revealed that children who had subtitled audio-visual input picked up more words than learners who did not have subtitles.

More recently, the effects of subtitled video have been compared with captioned video. Results revealed that captions lead to higher scores than subtitles on form recognition, form recall, and meaning recall tests of (low-) intermediate learners (e.g., Peters, 2019; Peters et al., 2016) after a one-off intervention. Captions also yielded better results than subtitles for promoting L2 learning of colloquial speech after watching 325 minutes of a TV series (Frumuselu et al., 2015). Other studies comparing subtitles and captions did not yield significant differences on beginning (Bisson et al., 2014) or low to intermediate L2 learners' meaning and form recall (Pujadas & Muñoz, 2019) or meaning recognition (Birulés-Muntané & Soto-Faraco, 2016) scores even after larger amounts of exposure (i.e., >110 min) to TV-programs (Fievez et al., 2020; Muñoz et al., 2021; Pujadas & Muñoz, 2019; Frumuselu, 2019).

While differences in terms of design and amount of exposure make it hard to compare findings across studies, results do not seem to highlight specific advantages of subtitles as compared with captions for L2 word learning. One possible explanation that may account for this finding is that subtitles do not facilitate speech stream segmentation and sound-script mapping, which may decrease learners' chances of noticing L2 forms and establishing form-meaning links.

4.5 Vocabulary learning from (captioned and subtitled) video and factors that influence learning

Even though there is a fair amount of research on vocabulary learning from audio-visual input, it is hard to determine and compare learning gains across studies because of the wide variation in terms of methodology and treatments. An analysis of the treatments used shows that they can be categorized in three groups based on the length of the treatment and the number of treatment sessions: (1) interventions based on short clips resulting in exposure of 9 to 20 minutes in single-session designs (e.g., Peters et al., 2016; Sydorenko, 2010); (2) interventions that have used longer clips (i.e., full-length episodes of TV programs) between 20 and 60 minutes in single-session designs (e.g., Birulés-Muntané &

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

Soto-Faraco, 2016; Peters & Webb, 2018); and (3) interventions that are based on multiple full-length episodes (e.g., Rodgers & Webb, 2020) in a multiple-session design (i.e., between 7 and 24 sessions). Learning gains found from pretest to posttest range from approximately 1 to 14 words learned[6] on a meaning recall or recognition test. However, these gains do not only depend on the presence of on-screen text, the duration of the treatment, and the number of sessions, but also on several other factors including word-related and learner-related variables.

Several properties of the target words – the words that are included in the pre- and posttests – may influence learning gains from watching (captioned or subtitled) audio-visual input. One such factor is cognateness, which has been defined as words from different languages that present an important overlap in form and meaning (e.g., the English word *monetary* and the Dutch word *monetair* in Peters & Webb, 2018). It has been found that cognateness has a facilitative effect on incidental vocabulary learning through watching (captioned or subtitled) audio-visual input (e.g., Fievez et al., 2020). In Peters and Webb (2018), for instance, the odds of a correct response on the meaning recall test were eight times higher when the target word was a cognate. These findings seem to indicate that there is a cognate facilitation effect in incidental vocabulary learning from video, which is consonant with other studies in the field of L2 vocabulary learning and processing (see Peters, 2020, for a comprehensive review).

Frequency of occurrence, that is, the number of instances of a word in the input, can have a facilitative effect on vocabulary learning from video as well. Repeated encounters have been found to positively influence the learning of single words through (captioned) audio-visual input in studies using a single viewing session (e.g., Peters & Webb, 2018) as well as after watching several videos (e.g., Rodgers & Webb, 2020). The positive effect of frequency is in line with studies on the role of frequency in written input but it remains unclear how many encounters are needed through audio-visual input in order to result in considerable learning. In studies using a multiple-session design (i.e., multiple viewing sessions), stronger effects of frequency on word learning have been reported when words appear repeatedly in a single episode – that is, through massed repetition – than when they appear spaced across episodes (e.g., Fievez et al., 2021; Rodgers & Webb, 2020). This is in line with findings from a meta-analysis (Uchihara et al., 2019) that revealed that massed conditions yielded larger effects than spaced conditions in incidental vocabulary learning research.[7]

Words that are accompanied by imagery (i.e., visual support provided by the images used in the video) may lead to higher scores on vocabulary posttests. Sydorenko's results (2010), for instance, showed that beginning learners had higher scores on a translation test for words that were depicted in the images than for words without visual support after watching short excerpts from a Russian comedy series. She therefore concluded that 'the meanings of some new words can be learned from very difficult authentic videos when the language is well-supported by visual images' (2010, p. 61). The supportive role of imagery for learning the meaning of unknown words was further confirmed by Peters (2019) who found that learners were three times more likely to be correct on a form recognition and meaning recall test when the meaning of target words was clarified in the images that occurred in a timeframe of five seconds before or after the use of the target in the auditory trace of the documentary (i.e., the spatial and temporal contiguity principle as presented in section 3.1).

Finally, word learning from watching audio-visual input is also affected by learner-related variables. There is a growing body of literature that has identified learners' prior vocabulary knowledge as an important variable in determining the amount of incidental learning that takes place while watching (captioned and subtitled) audio-visual input. More particularly, higher learning gains have been reported for learners with larger vocabulary sizes (e.g., Montero Perez, 2020).[8] This is often referred to as the Matthew effect or the 'rich get richer principle' as was already reported in early work on this topic (Baltova, 1999) in which it was also found that learners with a better oral fluency learned more vocabulary (Neuman & Koskinen, 1992). This has been explained by the fact that learners who know more words tend to have a better lexical coverage of the input, which may help learners to segment the speech stream and to notice unknown words in the input. Similarly, learners with larger scores on

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

general proficiency tests have also been found to obtain larger vocabulary learning gains after watching captioned or subtitled video than learners with lower scores (Muñoz et al., 2021).

In order to deal with the moderating role of word-related and learner-related factors, studies have recently shifted from analyses of (co)variance towards logistic regression and mixed effects models. The advantage of these types of analyses is that they are based on items and therefore allow to include different input-related variables such as the frequency of occurrence of words in the input, their visual representation, cognateness, as well as learner-related variables (e.g., L2 learner' proficiency level) in one model. This is valuable when target items have different properties as is typically the case in studies on audio-visual input because of the difficulty of manipulating target items in the input. However, it remains crucial to consider word-related factors when selecting target words.

5. Promoting grammatical knowledge, listening, and pronunciation through (captioned and subtitled) audio-visual input

5.1 Learning grammar and syntax through watching video with different types of on-screen text

The effects of audio-visual input and on-screen text on the learning of grammar have remained largely unexplored. Early work (e.g., d'Ydewalle & Van De Poel, 1999) on this topic found that children (8–12 years old) did not seem to show significant learning of morphology or syntax in the L2 (French or Danish) after watching a 10-minute video with subtitles or reversed subtitles (see Table 1 for a definition). Similarly, Van Lommel et al. (2006) investigated the acquisition of Esperanto grammar rules after a 16-minute subtitled video. Results indicated that watching the video without explicit instruction about grammar rules did not result in grammar uptake with scores around chance level. This finding is not surprising and might be explained by the nature of grammar learning that consists of the acquisition of grammatical regularities. To abstract these regularities incidentally from input, large amounts of exposure including repeated encounters with target constructions will be needed for learners to apply these rules to new utterances. This distinguishes grammar learning from vocabulary learning, which consists of learning items rather than regularities.

Research into the effects of more substantial exposure to L2 input on the learning of grammar and syntax is scarce. One exception is the study by Ghia (2012), who investigated whether watching five subtitled films could stimulate L2 learners' acquisition of syntax (i.e., the ordering of main constituents within phrases and clauses such as clefts, interrogative clauses, and complex phrases). More specifically, she studied the concept of 'translational salience', which refers to the relationship between the L2 audio and the translation of the audio in the subtitles. For instance, subtitles can present reductions or substitutions or literal transfer between L1 and L2. Subtitles can also include 'discrepancies in subtitle translation' (p. 90). Results indicated that the viewing activity resulted in initial learning of L2 syntax irrespective of the translation strategy used.

Pattemore and Muñoz (2020) investigated whether watching ten episodes of a TV-series (totaling 227 min) could stimulate L2 learners' acquisition of grammar. Participants were undergraduate EFL learners who were assigned to a captioned video or uncaptioned video condition. The authors found significant increases from pretest to posttest on the grammatical constructions test and students in the captioned group outperformed learners in the uncaptioned group. A significant interaction effect between working memory and the presence of captioning was also reported, which revealed that learners who did not have captions seemed to rely more on their working memory than learners who had captions, irrespective of their proficiency level. In addition, working memory – that is, the system that temporarily stores and processes information during higher order task performance – was linked to grammar learning gains in the uncaptioned group with better learning of grammatical constructions found in learners with greater working memory capacity. This is in line with Gass et al. (2019, p. 84), who argued that captions 'may neutralize some of working memory's limiting effects on language learning'. In a follow-up study using similar input conditions (Muñoz et al., 2021), it was found that frequency of occurrence of target grammatical constructions correlated with grammar

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

learning gains of upper-intermediate university L2 learners, especially for learners who did not have access to captions.

Drawing on the importance of noticing for L2 learning, captions with typographical enhancement have been examined. Lee and Révész (2020) found that captions with highlighted structures helped advanced EFL learners to outperform a captioning and a control group, even with treatment tasks that were based on 24 short clips of 20 to 50 seconds each. The use of captioned video with enhancement thus seems to encourage learners to reflect more about the target structures, hence resulting in learning gains. Similarly, Cintrón Valentín et al. (2019) investigated the effects of captions with/without highlighted target structures on L2 grammar learning. Crucially, participants watched multimodal input that was specifically created for this study and that consisted of animated videos in which the target items were carefully controlled for frequency of occurrence as well as for their placement in the video. They reported positive effects of textually enhanced captions on the learning of some of the selected target grammatical structures. The authors also argued that some target grammatical structures have structure-specific characteristics which may have mediated the effects of captioning and highlighting. This finding was further echoed in Cintrón-Valentín and Garciá-Amaya (2021).

5.2 Effects of (captioned and subtitled) audio-visual input on L2 listening and pronunciation

Being able to process the speech stream and automatically recognize word boundaries of L2 speech – that is, learners' bottom-up listening processes (Rost, 2002) – is essential for successful L2 listening comprehension. To promote L2 learners' bottom-up processes, studies have, for instance, used repeated or intensive listening activities in which learners process a particular input repeatedly and segment per segment (e.g., Matthews et al., 2014 on audio-based listening). Another approach consists of providing L2 learners with captioned video. In past research, captioned video had, however, been criticized because captions were expected to hinder listening (Diao et al., 2007, on audio-based listening) as well as the development of strategies needed to deal with authentic listening situations (Pujolà, 2002; Vandergrift, 2004). However, there are now first indications that captioned video stimulates speech perception (see Bird & Williams, 2002 for early work on this topic) in such a way that its effects are generalizable beyond the input used in the treatments. This was the case in Charles and Trenkic (2015), who report on a small-scale study investigating advanced L2 learners' speech segmentation ability after watching two 30-minute excerpts from two English documentaries in one of the three conditions (captioned video, sound only, or text only). It was found that participants in the captioned group showed most improvement on the speech segmentation test, which consisted of known sentences, but also unknown ones.

Other studies have examined the value of captioned video as compared with subtitled video for L2 speech decoding. Mitterer and McQueen (2009) found that advanced users of English could adapt to an unfamiliar accent after only 25 minutes of exposure to captioned video. In addition, captions were also found to lead to enhanced perceptual adaptation for both new and previously encountered words. Subtitles, however, were only helpful for known utterances and seemed to 'create lexical interference' (Mitterer & McQueen, 2009, p. 1), which hindered speech adaptation for new items. Furthermore, Birulés-Muntané and Soto-Faraco (2016) reported that watching a one-hour episode of an English TV drama with captions led to significantly more improvement of higher-intermediate English as a Second Language (ESL) learners' scores on a fill-gap listening test (i.e., learners had to fill in 24 words in a text while listening to a fragment) than watching video with subtitles or video only. However, they also pointed to a potential trade-off between subtitles and captions. Whereas the development of L2 speech perception was better with captions, scores on the plot comprehension test were highest with subtitles. Wisniewska and Mora (2020) found that an eight-week intervention program in which advanced learners watched 16 episodes of the British TV-series *Luther* with or without captions influenced L2 learners' speech processing skills. Importantly, the intervention program did not only stimulate learners' speech segmentation but also enabled them to process speech in a faster and

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

more efficient way, both of which are crucial in the development of listening fluency (Rost, 2014). Surprisingly, no differences between the captioned and video-only group were found. This might be owing to the advanced proficiency level of the participants or to the treatment, which was substantially longer than the treatment used in the abovementioned articles. Finally, Wisniewska and Mora found that captioned video led to pronunciation gains, as previously suggested by Mora and Cerviño-Povedano (2019). However, gains were also found in the uncaptioned video condition when learners' attention was directed to phonetic form.

6. L2 learners' use and perceptions of audio-visual input

6.1 L2 learners' perceptions towards the use of (captioned and subtitled) audio-visual input for L2 learning

Few studies have focused explicitly on perceptions and attitudes towards using video for L2 learning. There are, however, indications that learners find audio-visual input motivating and consider it a useful resource for L2 learning (e.g., Sockett & Kusyk, 2015; Wang, 2012). Regarding the use of on-screen text, it has been found that captioned video may result in more relaxed viewing (Pujadas & Muñoz, 2019) and may increase learners' self-confidence as language learners (Teng, 2021). In addition, captions do not seem to distract learners or hamper their enjoyment (Kruger et al., 2017). In one of the few studies based on qualitative data, Vanderplank (2019) investigated L2 learners' use of captioned video in an informal context. Participants watched video with captions over a period of six to 12 weeks. Importantly, they were in control of the viewing material. An analysis of viewing diaries revealed changes in attitudes and viewing behavior, resulting in three different usage groups: minimal users whose focus was on enjoying films, maximal users who used films for L2 learning, and evolving users who reported using films increasingly as a tool to learn languages. Vanderplank also argued that learners might need advice or guidance in order to maximally benefit from L2 audio-visual input.

In a large-scale study, Mariotti (2015) focused on L2 learners' use of audio-visual input with on-screen text. Participants were 1,374 secondary school and university students enrolled in formal language learning and 555 adult learners who watched video with on-screen text in an informal learning context. Participants who watched subtitled video[9] in a formal context thought that it had a positive effect on their listening and speaking skills. In addition, 70% of the students indicated that they watched more subtitled video in their spare time. However, students also indicated several problems such as the speed of the subtitles, which was often perceived as too fast. Of the participants who used video with on-screen text in an informal context, 60% found that it was time well spent and thought it helped them improve their listening skills. Most of the participants seemed to prefer subtitles (rather than captions or reversed subtitles) when watching for pleasure. In addition, learners who felt more confident also watched more subtitled video and reported to be more active in conversations. Finally, they also observed that learners who were very motivated to learn the L2 preferred captions instead of subtitles.

6.2 L2 learners' self-reported use of input modes in (captioned) video

Findings from self-reported data on L2 learners' use of input modes (i.e., audio, images, on-screen text) in captioned video (Sydorenko, 2010) indicated that beginning L2 learners primarily focused on the on-screen text, followed by the images and the audio. However, learners found the visuals more useful than the captions and the audio. In order to assess how much learners rely on captions (i.e., textual information), Leveridge and Yang (2013) developed a 40-item multiple-choice test including questions on sections where the captions match the audio (congruent items) and questions that focus on excerpts in which captions and audio do not match (incongruent items). The score gives an indication of learners' caption reliance and it was found that learners' score correlates with their

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

perceived captioning reliance (Leveridge & Yang, 2014). However, this information is based on self-reported data and it does not include information on learners' use of images.

Drawing on the concept of multimodality (see Section 2.1), studies have recently tried to tap into learners' use of a wide range of semiotic resources when watching audio-visual input. Compared with previous research in which a distinction was made between images, on-screen text, and audio, Ruck (2020) investigated L2 learners' self-reported foci of attention on 12 possible sources of input in the videos (i.e., an excerpt from a film and from a documentary) such as vocabulary, visuals, (non linguistic) sounds, cultural products (e.g., books, tools) or behaviors. She found that learners' self-reported use of these sources, as measured by means of a questionnaire, differed in both genres and that learners paid, for instance, more attention to linguistic features in the documentary than in the film clip. Abrams (2016) included a TV-series in technology-mediated task cycles for beginning L2 learners of German. She investigated the potential contribution of tasks to learners' semiotic awareness, that is, their awareness to use different aspects of multimodal input for meaning extraction. She found that the use of captioned audio-visual input contributed to learners' awareness of the connections between different semiotic resources such as, for instance, the link between actions and body language depicted in the input.

6.3 L2 learners' use and processing of captions and subtitles: Findings from eye-movement studies

An important addition to the literature on L2 audio-visual input since Vanderplank's (2010) article is the use of online data such as eye tracking in order to understand learners' use and processing of input modalities (see Doherty & Kruger, 2018 and Kruger et al., 2015 for reviews on eye tracking in subtitling research).[10] Eye-movement technology allows for relatively unobtrusive data collection and makes it possible to register eye movements while participants engage in a natural viewing activity (see Godfroid, 2020 and Conklin et al., 2018 for methodological discussions on eye-tracking and multimodal input). It should be noted that eye-tracking studies with audio-visual input are relatively scarce and have focused on two main areas of interest, that is, learners' time spent on the input modalities (i.e., on-screen text and images) and learners' time spent on specific words or structures in the on-screen text. So far, however, the question of how the dynamic nature of the input (i.e., moving images with moving text that is synchronized with audio) impacts learners' reading processes remains largely unanswered.[11]

With regard to the amount of time spent on the captioning or subtitling area, Bisson et al. (2014) reported reading percentages of 43% for the captions and 51% for subtitles. Even though learners were not familiar with the language used in the captions (i.e., Dutch), findings reveal that they spent a significant amount of time on the on-screen text. The amount of time spent on the captions in Winke et al. (2013) was slightly higher than in Bisson et al. (2014), with 66% for familiar content and 69% for unfamiliar content. Differences in script may, however, affect eye-movements. For instance, L2 learners of Arabic spent more time reading the captions than L2 learners of Spanish and Russian. Fixation times of L2 learners of Chinese presented a lot of variation and they varied in function of content familiarity (74% for unfamiliar vs. 62% for familiar content). In a study with different age and proficiency groups, Muñoz (2017) reported that children (who were mainly beginning EFL learners) tended to skip subtitles less and had a longer total fixation duration on the subtitle area than older learners. More fixations were found on captions than on subtitles in children's fixation data, which may indicate that captions are more challenging to process for this age and proficiency group. Older and higher-level learners showed a more regular reading pattern of the captions but tended to skip subtitles more often. Finally, working memory has also been found to play a role in L2 English learners' time spent on captions (Gass et al., 2019). Findings indicated, for instance, that learners with a higher working memory spent less time rereading the captions during a second viewing.

Another question that has been addressed is whether learners are able to simultaneously deal with on-screen text and imagery, which are both visual input sources. Fixation times from eye-movement

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

analyses have shown that learners are able to process the images even when subtitles or captions are shown and read (e.g., Bisson et al., 2014). This is in line with earlier work that had shown that learners are able to shift their attention between images and subtitles effortlessly and automatically (d'Ydewalle & Gielen, 1992) while watching a subtitled program with two-line subtitles (d'Ydewalle & De Bruycker, 2007). Following up on d'Ydewalle and De Bruycker (2007), Perego et al. (2010) found similar performance on a subtitle recognition test (i.e., multiple choice questions on the exact wording in the subtitle) and scene recognition task (i.e., 'Did you see the scene in the video?') and reported that scene recognition was still successful even if learners spent less than 40% of their fixations on the visuals. According to these data, L2 learners can process and integrate information from multiple competing sources of information. These findings also point to the cognitive effectiveness of subtitle and image processing and suggest that there does not seem to be a trade-off between the processing of images and subtitles.

Eye-tracking has also been used to quantify learners' time spent on specific words or structures in on-screen text as a function of attention-drawing techniques. Regarding the effect of textual enhancement through highlighting, it was found that highlighting words in captions (Lee & Révész, 2020) was successful in directing learners' attention to the target grammatical constructions with significantly more rereading times and visits to highlighted target structures. Higher rereading times were also found by Ghia (2012) who used translational salience (i.e., including a contrast between the L2 structure and the L1 translation in the L1 subtitle) as attention-drawing technique. Eye-movement analyses in Montero Perez et al. (2015) revealed an interplay between two attention-drawing techniques: visual salience (full vs. keyword captions) and test announcement (i.e., informing vs. not informing learners about upcoming vocabulary tests).

Finally, it has been asked whether learners' fixations on target items are related to learning gains. In order to do so, the relationship between eye-movement data (i.e., online measures) and offline measures (i.e., learners' scores on posttests) have been considered (see Godfroid et al., 2013 on incidental vocabulary learning while reading). For instance, Lee and Révész (2020) found that attention to the target structures in the captions with highlights was linked to greater performance on a written and oral productive test but only for one of the target grammatical structures (i.e., present perfect). Montero Perez et al. (2015) found that eye-movement data of learners in the captioning incidental learning group significantly correlated with learning. For instance, learners' first pass time (i.e., the amount of time spent on a word before the eyes leave that word) on a given word correlated positively with L2 learners' form recognition of that word, indicating that longer first pass time increases the likelihood of correctly recognizing a word on the posttest.

6.4 L2 learners' exposure to audio-visual input outside of the L2 classroom and its effect on L2 learning

Recently, studies on informal language learning or language learning beyond the classroom (see Reinders & Benson, 2017 on dimensions and definitions of these concepts) have identified audio-visual input as an important source of L2 input in out-of-school contexts. This is not surprising since audio-visual input is ubiquitous and can be easily encountered in daily life. It has, for instance, been shown that watching audio-visual input (in the L1 or the L2) is one of the most common activities in everyday life (European Commission, 2019) with 81% of the people who filled out the survey watching TV (almost) every day.

It has been reported that L2 learners of English frequently engage with audio-visual input in the target language in out-of-school contexts (e.g., Lindgren & Muñoz, 2013; Muñoz et al., 2018; Verspoor et al., 2011). Peters (2018) found that more than 40% of the surveyed participants (i.e., Dutch-speaking adolescent L2 learners of English in Flanders) regularly watched English TV (with or without subtitles). Crucially, watching (subtitled) TV in out-of-school contexts was related to aspects of L2 development such as L2 learners' receptive vocabulary knowledge. This was also the case in follow-up research in Flanders that found that watching TV promotes receptive (Puimège &

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

Peters, 2019a) and productive vocabulary knowledge (Bollansée et al., 2021) of L2 learners of English. González Fernández and Schmitt (2015) found that watching films, television, and video correlated with knowledge of certain collocations in English in adult Spanish learners of English. In addition, Lindgren and Muñoz (2013) found that watching subtitled films was not only one of the most common activities of primary school EFL learners in seven European countries, but it also had a strong impact on learners' listening and reading skills. Similarly, De Wilde et al. (2020) showed that watching subtitled television correlated positively with L2 knowledge of 10–12-year-old children who did not yet have formal L2 instruction. However, it was not as important as social media, gaming, and speaking English in this particular age group. Kusyk and Sockett (2012) reported that half of the students in their participant pool watched English television weekly by downloading or streaming TV series. Interestingly, it was found that frequent TV viewers had significantly more knowledge of frequently occurring multiword combinations than non-frequent viewers of English TV.

It should, however, be stressed that the abovementioned studies focused on L2 English; little is known about the use and effects of audio-visual input in out-of-school contexts for other target languages. One exception is the article by Peters et al. (2019), in which they compared the impact of out-of-class exposure on Flemish L2 learners' vocabulary knowledge in English and French. It was found that L2 learners' exposure to French was generally limited to formal classroom-based instruction while they engage daily with (spoken) English language input through songs, TV, internet, and gaming. This was found to have an impact on participants' vocabulary knowledge with significantly lower scores for French than for English. Drawing on the impact of input exposure in out-of-school-contexts, teachers should be encouraged to prepare and stimulate language learners to engage in out-of-class viewing and listening activities, especially for contexts where the L2 is not present in daily life and where exposure is limited to the L2 classroom, the teacher, and the textbook (Tragant et al., 2016).

Finally, since these viewing activities take place in out-of-class contexts, they may be particularly beneficial for autonomous or self-directed learners who can fully exploit or regulate their L2 viewing experiences. In this respect, Cole (2015; as referenced in Vanderplank, 2020) argued that autonomous learners were more focused on the linguistic elements of English language TV programmes than class-based learners.

7. From research to practice

7.1 Use of audio-visual input in the classroom

King (2002, p. 510) stated that audio-visual input 'compensates for many of the shortcomings in the EFL experience by bringing language to life'. It has also been argued that video provides a kind of authentic and naturalistic exposure to spoken language that could not be realized by textbooks alone (Bednarek, 2018; Jones & Horák, 2014). However, little is known about the role of audio-visual input as part of a classroom activity or as part of a specific didactical approach.

Recently, audio-visual input has been used to operationalize input-based tasks[12] that are used to provide a rich source of input in a task-based language teaching (TBLT) approach. In Lee and Révész (2020), for instance, short news clips of 20 to 50 seconds long were used that contained two target grammatical constructions (i.e., verbs in present perfect and past simple) and were either presented without captions, with captions, or with enhanced captions (i.e., the target construction is highlighted in the caption). Next, Korean ESL learners were asked to play the role of a newsroom editor. This meant that they had to watch the clips and decide, for each of the clips, whether the title and news category were appropriate. Next, they had to watch news clips in Korean and report to their friends what they saw in English in an oral and written production test.

So far, large-scale systematic investigations into the use of video in course materials and in teachers' day-to-day practices are lacking. One exception is the chapter of Mariotti (2015) that discusses the use of video with on-screen text (i.e., captions, subtitles, reversed subtitles) in L2

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

teaching by 45 European foreign language teachers. She found that 39 out of the 45 surveyed teachers were familiar with audio-visual materials for L2 teaching. About one third of the surveyed participants said they used subtitled materials[13] a lot while 25.6% used them sometimes. Teachers' main reason for using subtitled audio-visual input is to improve students' oral comprehension. They also considered audio-visual materials useful for increasing L2 learners' motivation as well as their listening skills and vocabulary learning. In addition, Mariotti's findings revealed that teachers considered subtitles appropriate for beginning learners whereas captions were believed to be more appropriate for intermediate and advanced learners. Teachers also reported they used subtitled audio-visual materials for introducing L2 learners to course contents, to expose them to the L2 or in L2 tests. However, they also mentioned that they need support on how to use on-screen text effectively in the classroom.

Finally, there are also indications that TV-series are 'underutilized' in ESL Classrooms, as was argued by Bednarek (2018, p. 224) based on L2 learners' responses to questions such as 'my English teacher has used/shown examples from English-language TV series in the classroom'. Similarly, Vanderplank (2016a) stated that the implementation of captioned video viewing in the classroom is still limited even though it is recognized as a valuable activity in informal out-of-school settings.

7.2 Choosing a type of on-screen text for viewing activities

This article has discussed the potential of audio-visual input for fostering different aspects of L2 learning. In addition, it has reviewed findings on the role of different types of on-screen text for supporting L2 learners and promoting L2 learning gains. Table 3 presents a tentative summary of research findings to schematically represent the potential language learning benefits that are associated with the different types of on-screen text. Two main findings seem to emerge from this table. First, different types of on-screen text may lead to different types of learning gains but most robust results have been found for captioned video. Second, most studies have looked into the potential of on-screen text for vocabulary learning.

While on-screen text may present numerous linguistic benefits, not all videos come with options to activate on-screen text. If teachers want to use on-screen text or give learners the possibility to activate on-screen text, this often means that they have to create materials themselves. However, this requires not only technical training (i.e., to use subtitling software and video editing tools) but also represents a serious time investment, especially when the input still needs to be transcribed/translated before on-screen text can be added. Luckily, technological advances and streaming services have made different types of on-screen text (e.g., captions, glossed captions) more easily available, which may contribute to the more widespread use of these types of materials inside and outside the classroom. We also find several dedicated evidence informed L2 learning tools. IdiomsTube (Lin, 2021), for instance, is a computer-assisted learning tool that aims at promoting formulaic language learning from English YouTube videos[14] with captions. To do so, it automatically assesses video difficulty, generates pre-learning activities and tasks, and recommends videos. Another example is the 'Language Reactor' that offers different tools to promote L2 learning from authentic materials such as bilingual subtitles, captions with glosses, dictionaries, and so forth, for a wide range of Netflix series and YouTube videos, including different target languages.

To further transfer insights on the role of video and on-screen text from research to teaching practice, at least two aspects should be considered. The first consists in sharing research results with L2 teachers through practical sessions that may also involve technical training and worked-out scenarios with audio-visual input for L2 teaching. Second, it should be noted that publishers of learning materials (e.g., textbook publishers) can play an important role in selecting and adapting different types of audio-visual materials for classroom use to familiarize learners from the start with audio-visual input and different types of on-screen text.

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

Table 3. Potential language learning benefits associated with the use of different types of on-screen text as compared to video only (without on-screen text)

	Detrimental effect	Mixed findings	Facilitative effect
Vocabulary		• Subtitles • Keyword captions • Captions with typographic enhancement	• Captions • Captions with glosses
Comprehension		• Keyword captions • Bilingual subtitles	• Subtitles • Captions
Listening	Subtitles		Captions
Pronunciation		Captions	
Grammar		• Captions with typographic enhancement • Subtitles	Captions

7.3 Preparing learners in the classroom for out-of-class viewing

Webb (2015) argued that one of the greatest benefits of audio-visual input is that it can provide large amounts of exposure to spoken L2 input because of its overall availability and accessibility. Audio-visual input therefore holds great potential for extensive viewing or the 'regular silent uninterrupted viewing of L2 television inside and outside of the classroom' (Webb, 2015). The concept of extensive viewing is based on previous research into the role of extensive reading programs (e.g., Cobb, 2007) that had revealed positive effects on, amongst others, comprehension (e.g., Nakanishi, 2015) and vocabulary learning (e.g., Webb & Chang, 2015).

To promote regular out-of-class viewing, initial classroom-based viewing is required (Webb, 2015). L2 teachers thus play a crucial role in promoting learners' use of audio-visual input and in establishing 'a dialogue with learners about their viewing and listening habits' (Sockett, 2012). In addition, they need to raise learners' awareness (Webb, 2015) of the benefits of L2 audio-visual input and prepare them in the classroom for autonomous viewing. So far, research that has explicitly focused on in-classroom preparation for extensive viewing is lacking, but several questions could be addressed, including the following: How can we train teachers and learners in effective use of audio-visual input? What is the most effective way to promote use of audio-visual input in out-of-class contexts?

Even though research on this topic is lacking, there are at least three recommendations that can be formulated to prepare learners for out-of-class viewing activities.

Recommendation 1: Select appropriate material in function of the specific objective of the viewing activity (e.g., listening, comprehension, grammar, vocabulary).

The selection of appropriate materials for L2 learners is a difficult task.[15] However, this review has put forward several criteria that influence video comprehension as well as learning gains. These findings suggest that a careful selection of videos that is based on various parameters such as the lexical coverage (see section 2.3), the potential contribution of imagery, and the role that on-screen text may play for stimulating different learning goals (see Table 3 for a summary) can lead to significant learning. For instance, teachers might want to consider using captions to promote learners' speech decoding and comprehension as well as to promote incidental vocabulary learning.

Recommendation 2: Raise learners' awareness of the benefits of video by showing how it can support L2 learning.

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

This review has discussed different language learning benefits (e.g., vocabulary) that have been found after watching audio-visual input. While some of these benefits were found after relatively short exposure to audio-visual input, more substantial learning gains can occur when learners engage more extensively with video in out-of-classroom contexts. By drawing learners' attention to gains that can occur through watching video in the classroom, we may raise learners' awareness of the potential of video material for L2 learning, which is crucial if we want to stimulate learners to engage in independent out-of-class viewing.

Recommendation 3: Raise learners' awareness of different types of on-screen text by using different options in the classroom and by explaining how they can support L2 learning.

To educate learners about the support that on-screen text can offer, we could teach them how to use these options. We could also highlight the advantages and drawbacks that have been associated with different types of on-screen text through different types of viewing activities in the classroom. For instance, subtitles might be helpful for beginning L2 learners to help increase their comprehension of authentic L2 input.

8. Concluding remarks and future directions

This article presents a summary of the main topics and findings that can be distilled from the considerable number of studies on L2 audio-visual input that have been published since 2010. Not all topics have received similar attention and studies have often used very different designs, which makes it hard to draw firm conclusions or formulate clear-cut recommendations. Despite this, it may be asked how the field has evolved in the past decade. At least four aspects can be considered. A first evolution is related to the more widespread use of L2 audio-visual input in general. The growing number of studies that have highlighted its potential in both formal and informal contexts have shown that such input is a rich source, which can bring natural language use and authenticity in formal contexts or create rich L2 learning experiences in out-of-class contexts. This can, for instance, be seen in studies that have revealed a positive relation between exposure to audio-visual input and aspects of L2 proficiency (see Section 6.4). Second, numerous studies have investigated the role of audio-visual input in relation to incidental (vocabulary) learning and have shown how a complex interplay of mechanisms including the role of on-screen text, imagery, word-related variables (e.g., cognateness), and learner-related variables (e.g., prior vocabulary knowledge) influence incidental L2 learning from audio-visual input. Third, multiple-session treatments have been used to increase the duration of video exposure (e.g., watching an entire season of a TV-show). These studies refine our understanding about the potential effects of watching video on L2 learning over time as well as the contribution of on-screen text in extensive viewing settings. A final contribution that holds great potential for further research is related to the use of new types of data collection techniques such as eye-tracking, tracking-and-logging data (e.g., tracking learners' use of glossed captions), and measurement instruments that consider the multimodal nature of the input (e.g., tests that include the role of imagery). These techniques can further elucidate how learners interact with the different resources that can be found in audio-visual input.

When considering the generalizability of the findings presented in this review, several questions for future research emerge. First, the bulk of research into audio-visual input for L2 learning has been conducted in ESL and EFL settings. Studies in which the effects of audio-visual input for the learning of other target languages have been studied are scarce (e.g., Montero Perez, 2020 on French; Sydorenko, 2010 on Russian) and few studies have compared gains in different target languages (see Winke et al., 2010 for an exception).

Second, studies have targeted learners with different backgrounds (e.g., secondary school pupils, primary school learners, university learners) and different proficiency levels. However, it remains difficult to extrapolate findings to all groups of learners. For instance, it is unclear whether beginning learners can process captions and how that affects their auditory processing. There are findings for different age groups, but it remains to be seen how younger learners (<10) deal with L2 audio-visual

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

input and on-screen text. In addition, individual differences have been included in some studies and aspects such as working memory and prior vocabulary knowledge seem to play a role in the effects of audio-visual input on L2 learning. However, it remains unclear whether and to what extent L2 learners, irrespective of age, educational background, and literacy can benefit from audio-visual input.

In terms of linguistic benefits, this review has shown that a considerable number of studies have focused on L2 vocabulary learning. There are indications that audio-visual input has beneficial effects on other aspects of language development such as grammar learning and pronunciation but this needs further investigation. Future research might also consider the use of audio-visual input to stimulate learners' development of culture (Sokoli & Zabalbeascoa, 2019), media, and television literacy (Bednarek, 2018).

Research that capitalizes on the affordances of technology is scarce (see Mirzaei et al., 2017 and Lin, 2021 for exceptions). However, a combination of automatic speech recognition and natural language processing techniques could play an important role in making (automatically generated) captions and glossed captions more easily available. Technology also plays an important role in obtaining detailed learning analytics on learners' use and performance in video-based platforms. This might be an important next step to gauge L2 learners' use of audio-visual input as well as its effectiveness from a big data perspective in informal contexts (see Sauro & Zourou, 2019 on the digital wilds). Similarly, detailed learning analytics could be used to personalize audio-visual input for L2 learners (e.g., highlight words that will be difficult for a learner with a particular proficiency level).

Finally, the potential of audio-visual input in informal and formal learning contexts should be further promoted in the L2 classroom if we want to stimulate learners from across the globe to take advantage of the many opportunities that this type of input offers for L2 learning. In addition, technological affordances will continue to increase the possibilities of audio-visual input for L2 learners. Consequently, there is no doubt that audio-visual input will play an important role in future L2 research as well as in the L2 classroom and beyond.

Questions arising

1. How could audio-visual input be implemented effectively in L2 teaching, and which insights would teachers and learners need to do so?
2. How could L2 teachers motivate their learners to engage in out-of-class viewing activities?
3. What challenges are related to the use of audio-visual input for L2 learning and teaching, and how could research deal with these challenges?
4. What other linguistic benefits (i.e., not studied in previous research) for L2 learning could be promoted through watching L2 audio-visual input?
5. What are methodological weaknesses in research on audio-visual input and how could we tackle these?
6. Which factors should we try to control for in the design of the learning materials used in research on audio-visual input, and how should future research take into account the role of moderating variables (e.g., input-related factors and learner-related variables)?

Acknowledgements. I am grateful to the reviewers whose detailed and insightful feedback has greatly improved this article. I would also like to thank Tetyana Sydorenko and Lizz Huntley for their comments on an earlier draft.

Endnotes

[1] Anecdotal reports on autonomous language learning from watching video are often found in press and media, such as the following article in the *New York Times*: https://www.nytimes.com/2021/05/29/arts/television/friends-reunion-english.html.

[2] Dubbed video or other types of video that are not based on L2 auditory input are not considered in this review.

[3] Some studies have asked whether captioning can be used for all levels (e.g., Lwo & Chia-Tzu Lin, 2012; Taylor, 2005). Lavaur and Bairstow (2011) found that advanced L2 learners performed better after watching video without captions than after watching captioned video. Intermediate learners had similar comprehension scores in the captioning and video only group. While this study suggests that captioning effectiveness is influenced by proficiency, this seems to be an isolated finding.

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

In addition, one of the methodological difficulties when studying the role of L2 proficiency is that the same video excerpt was used for all proficiency levels while it has been argued that captions cannot overcompensate when the gap between L2 learners' knowledge and the knowledge needed for adequate comprehension is too large.

[4] It should be noted that few studies have exclusively focused on audio-visual input (without on-screen text). Numerous studies into the use of captions and subtitles have however included a video only group as comparison group.

[5] Incidental learning refers to learning gains that occur as a by-product of a meaning-focused activity (see Webb, 2020 for a discussion on the concept of incidental vocabulary learning), that is, watching audio-visual input. Participants in incidental learning studies are typically not informed about vocabulary posttests (see Hulstijn, 2003) but watch the input for meaning or in function of a comprehension task.

[6] These findings do not include studies in which explicit teaching was used (e.g., Suárez & Gesa, 2019).

[7] However, there is strong evidence that spaced or distributed practice is important in explicit L2 learning (e.g., Bird, 2010).

[8] However, the effects of learners' prior vocabulary knowledge seem to be less strong in studies with a multiple-session design (e.g., Rodgers & Webb, 2020) but more research is needed to refine our understanding of this particular aspect.

[9] It is not clear from the text if the author refers to subtitles, captions, or reversed subtitles. All three types of on-screen text were included in the study and the concept 'subtitles' seems to refer to all three of them.

[10] For early eye-tracking research into captioned and subtitled video, see, for instance d'Ydewalle et al. (1991) and d'Ydewalle and Gielen (1992).

[11] Few studies have addressed this question from a methodological point of view. One exception is Kruger and Steyn (2014), who report on the development and validation of a reading index to measure learners' reading behavior of dynamic texts over longer segments such as captions and subtitles. More particularly, it calculates the degree to which each subtitle was read by a particular participant.

[12] So far, input-based tasks have mainly been conceptualized as unimodal tasks that are based on written or auditory input.

[13] It is unclear from the study if 'subtitles' refers to all types of on-screen text studied (captions, subtitles, reversed subtitles) or to a specific type of on-screen text.

[14] Another example based on YouTube videos is Youglish.com that uses captioned YouTube videos to foster pronunciation learning in English.

[15] Rodgers (2016) argues that television input meets the conditions outlined by Nation (2007) to be suitable for L2 learning: it is available in large quantities, it is familiar to language learners, the input can foster vocabulary learning, and learners find the input interesting.

References

Abrams, Z. I. (2016). Possibilities and challenges of learning German in a multimodal environment: A case study. *ReCALL*, *28*(3), 343–363. https://doi.org/10.1017/S0958344016000082

Adesope, O. O., & Nesbit, J. C. (2012). Verbal redundancy in multimedia learning environments: A meta-analysis. *Journal of Educational Psychology*, *104*(1), 250–263. https://doi.org/10.1037/a0026147

Al-Surmi, M. (2012). Authenticity and TV shows: A multidimensional analysis perspective. *TESOL Quarterly*, *46*(4), 671–694. https://doi.org/10.1002/tesq.33

Arndt, H. L., & Woore, R. (2018). Vocabulary learning from watching YouTube videos and reading blog posts. *Language Learning & Technology*, *22*(3), 124–142. https://doi.org/10125/44660

Baltova, I. (1999). Multisensory language teaching in a multidimensional curriculum: The use of authentic Bimodal video in core French. *The Canadian Modern Language Review*, *56*(1), 31–48. https://doi.org/10.3138/cmlr.56.1.31

Bednarek, M. (2018). *Language and television series. A linguistic approach to TV dialogue*. Cambridge University Press.

Bianchi, F., & Ciabattoni, T. (2008). Captions and subtitles in EFL learning: An investigative study in a comprehensive computer environment. In A. Baldry, M. Pavesi, & C. Taylor Torsello (Eds.), *From Didactas to Ecolingua* (pp. 69–80). Edizioni Università di Trieste.

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied PsychoLinguistics*, *31*(4), 635–650. http://doi.org/10.1017/S0142716410000172

Bird, S. A., & Williams, J. N. (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. http://doi.org/10.1017/S0142716402004022

Birulés-Muntané, J., & Soto-Faraco, S. (2016). Watching subtitled films can help learning foreign languages. *PLoS One*, *11*(6), 1–10. https://doi.org/10.1371/journal.pone.0158409

Bisson, M.-J., van Heuven, W. J. B., Conklin, K., & Tunney, R. J. (2014). Processing of native and foreign language subtitles in films: An eye tracking study. *Applied Psycholinguistics*, *35*(2), 1–20. https://doi.org/10.1017/S0142716412000434

Black, S. (2021). The potential benefits of subtitles for enhancing language acquisition and literacy in children: An integrative review of experimental research. *Translation, Cognition & Behavior*. [Advance online publication]. https://doi.org/10.1075/tcb.00051.bla

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

Boers, F. (2021). Glossing and vocabulary learning. *Language Teaching*. [Advance online publication] https://doi.org/10.1017/S0261444821000252

Bollansée, L., Puimège, E., & Peters, E. (2021). 'Watch out! Behind you is the enemy!' An exploratory study into the relationship between extramural English and productive vocabulary knowledge. In V. Werner & F. Tegge (Eds.), *Pop culture in language education* (pp. 199–214). Routledge.

Borrás, I., & Lafayette, R. C. (1994). Effects of multimedia courseware subtitling on the speaking performance of college students of French. *Modern Language Journal*, *78*(1), 61–75. https://doi.org/10.1111/j.1540-4781.1994.tb02015.x

Caruana, S. (2021). An overview of audiovisual input as a means for foreign language acquisition in different contexts. *Language and Speech*, *64*(4), 1018–1036. https://doi.org/10.1177/0023830920985897

Charles, T. J., & Trenkic, D. (2015). Speech segmentation in a second language: The role of bi-modal 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*, 25–56. https://doi.org/10.1007/s42321-018-0002-8

Cintrón-Valentín, M. C., & Garciá-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

Cintrón Valentín, M. C., García-Amaya, L., & Ellis, N. C. (2019). Captioning and grammar learning in the L2 Spanish classroom. *The Language Learning Journal*, *47*(4), 439–459. https://doi.org/10.1080/09571736.2019.1615978

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology*, *11*(3), 38–63. http://dx.doi.org/10125/44117

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

Cross, J. (2011). Comprehending news videotexts: The influence of the visual content. *Language Learning & Technology*, *15*(2), 44–68. http://dx.doi.org/10125/44251

Davies, M. (2021). The TV and movies corpora. Design, Construction, and use. *International Journal of Corpus Linguistics*, *26*(1), 10–37. https://doi.org/10.1075/ijcl.00035.dav

deHaan, J., Reed, W. M., & Kuwada, K. (2010). The effect of interactivity with a music video game on second language vocabulary recall. *Language Learning & Technology*, *14*(2), 74–94. http://dx.doi.org/10125/44215

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2020). Learning English through out-of-school exposure. Which levels of language proficiency are attained and which types of input are important? *Bilingualism*, *23*(1), 171–185. https://doi.org/10.1017/S1366728918001062

Diao, Y., Chandler, P., & Sweller, J. (2007). The effect of written text on comprehension of spoken English as a foreign language. *The American Journal of Psychology*, *120*(2), 237–261. https://doi.org/10.2307/20445397

Doherty, S., & Kruger, J. L. (2018). The development of eye tracking in empirical research on subtitling and captioning. In T. Dwyer, C. Perkins, S. Redmond, & J. Sita (Eds.), *Seeing into screens: Eye tracking and the moving image* (pp. 46–64). Bloomsbury.

Durbahn, M., Rodgers, M., & Peters, E. (2020). The relationship between vocabulary and viewing comprehension. *System*, *88*, 102166. https://doi.org/10.1016/j.system.2019.102166

d'Ydewalle, G., & De Bruycker, W. (2007). Eye movements of children and adults while reading television subtitles. *European Psychologist*, *12*(3), 196–205. https://doi.org/10.1027/1016-9040.12.3.196

d'Ydewalle, G., & Gielen, I. (1992). Attention allocation with overlapping sound, image and text. In K. Rayner (Ed.), *Eye movements and visual cognition* (pp. 415–427). Springer-Verlag.

d'Ydewalle, G., Praet, C., Verfaillie, K., & Van Rensbergen, J. (1991). Watching subtitled television: Automatic reading behavior. *Communication Research*, *18*, 650–666. https://doi.org/10.1177/009365091018005005

d'Ydewalle, G., & Van de Poel, M. (1999). Incidental foreign-language acquisition by children watching subtitled television programs. *Journal of Psycholinguistic Research*, *28*(3), 227–245.

Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2018). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, *40*(2), 341–366. https://doi.org/10.1017/S0272263117000109

European Commission. (2019). *Media use in the European Union*. Retrieved from https://op.europa.eu/en/publication-detail/-/publication/c2fb9fad-db78-11ea-adf7-01aa75ed71a1/language-en

Feng, Y., & Webb, S. (2020). Learning vocabulary through reading, listening, and viewing. *Studies in Second Language Acquisition*, *42*(3), 499–523. https://doi.org/10.1017/S0272263119000494

Fievez, I., Montero Perez, 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. (2021). Promoting incidental vocabulary learning through watching a French Netflix series with glossed captions. *Computer Assisted Language Learning*, [Advance online publication]. https://doi.org/10.1558/cj.39370

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

Frumuselu, A. D. (2018). The implications of cognitive load theory and exposure to subtitles in English foreign language (EFL). *Translation and Translanguaging in Multilingual Contexts*, *4*(1), 55–76. https://doi.org/10.1075/ttmc.00004.fru

Frumuselu, A. D. (2019). A friend in need is a film indeed: Teaching colloquial expressions with subtitled television series. In C. Herrero & I. Vanderschelden (Eds.), *Using film and media in the language classroom* (pp. 92–107). Multilingual Matters.

Frumuselu, A. D., De Maeyer, S., Donche, V., & Gutiérrez-Colon Plana, M. (2015). Television series inside the EFL classroom: Bridging the gap between teaching and learning informal language through subtitles. *Linguistics and Education*, *32B*, 107–117. https://doi.org/10.1016/j.linged.2015.10.001

Gambier, Y., Caimi, A., & Mariotti, C. (2015). *Subtitles and language learning. Principles, strategies and practical experiences*. Peter Lang.

Gass, S., Winke, P., Isbell, D. R., & Ahn, J. (2019). How captions help people learn languages: A working-memory, eye-tracking study. *Language Learning & Technology*, *23*(2), 84–104. https://doi.org/10125/44684

Ghia, E. (2012). *Subtitling matters. New perspectives on subtitling and foreign language learning*. Peter Lang.

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

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words. *Studies in Second Language Acquisition*, *35*(03), 483–517. https://doi.org/10.1017/S0272263113000119

Goldstein, B., & Driver, P. (2015). *Language learning with digital video*. Cambridge University Press.

González Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have? The effects of frequency and amount of exposure. *ITL, International Journal of Applied Linguistics*, *166*(1), 94–126. https://doi.org/10.1075/itl.166.1.03fer

Guillory, H. (1998). The effects of keyword captions to authentic French video on learner comprehension. *CALICO Journal*, *15*(1), 89–108. http://dx.doi.org/10.1558/cj.v15i1-3.89-108

Herrero, C., & Vanderschelden, I. (2019). *Using film and media in the language classroom*. Multilingual Matters.

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

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in A Foreign Language*, *13*(1), 403–430.

Hulstijn, J. H. (2003). Incidental and intentional learning. In C. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 349–381). Blackwell Publishing Inc.

Jones, C., & Horák, T. (2014). Leave it out! The use of soap operas as models of spoken discourse in the ELT classroom. *Journal of Language Teaching and Learning*, *4*(1), 1–14.

Kato, F., Spring, R., & Mori, C. (2016). Mutually beneficial foreign language learning: Creating meaningful interactions through video-synchronous computer-mediated communication. *Foreign Language Annals*, *49*(2), 355–366. https://doi.org/10.1111/flan.1219

King, J. (2002). Using DVD feature films in the EFL classroom. *Computer Assisted Language Learning*, *15*(5), 509–523. https://doi.org/10.1076/call.15.5.509.13468

Koolstra, C. M., & Beentjes, J. W. 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

Kress, G., & Van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. Arnold.

Kruger, J., Doherty, S., & Soto-Sanfiel, M. (2017). Original language subtitles: Their effects on the native and foreign viewer. *Comunicar*, *1*, 23–32. https://doi.org/10.3916/C50-2017-02

Kruger, J., Szarkowska, A., & Krejtz, I. (2015). Subtitling on the moving image: An overview of eye-tracking studies. *Refractory: A Journal of Entertainment Media*, *25*, 1–14.

Kruger, J. L., & Steyn, F. (2014). Subtitles and eye tracking: Reading and performance. *Reading Research Quarterly*, *49*(1), 105–120. http://doi.org/10.1002/rrq.59

Kusyk, M., & Sockett, G. (2012). From informal resource usage to incidental language acquisition: Language uptake from online television viewing in English. *ASp, la revue du GERAS*, *62*, 45–65. https://doi.org/10.4000/asp.3104

Laaser, W., & Toloza, E. A. (2017). The changing role of the educational video in higher distance education. *International Review of Research in Open and Distance Learning*, *18*(2), 264–276. https://doi.org/10.19173/irrodl.v18i2.3067

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size, and reading comprehension. *22*(1), 15–30.

Lavaur, J.-M., & Bairstow, D. (2011). Languages on the screen: Is film comprehension related to the viewers' fluency level and to the language in the subtitles? *International Journal of Psychology*, *46*(6), 455–462. doi: 10.1080/00207594.2011.565343.

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), 1–27. https://doi.org/10.1017/S0272263120000108

Lee, P.-J., Liu, Y.-T., & Tseng, W.-T. (2021). One size fits all? In search of the desirable caption display for second language learners with different caption reliance in listening comprehension. *Language Teaching Research*, *25*(3), 400–430. https://doi.org/10.1177/1362168819856451

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

Leveridge, A. N., & Yang, J. C. (2013). Testing learner reliance on caption supports in second language listening comprehension multimedia environments. *ReCALL*, *25*(2), 199–214. http://doi.org/10.1017/S0958344013000074

Leveridge, A. N., & Yang, J. C. (2014). Learner perceptions of reliance on captions in EFL multimedia listening comprehension. *Computer Assisted Language Learning*, *27*(6), 545–559. https://doi.org/10.1080/09588221.2013.776968

Liao, S., Kruger, J. L., & Doherty, S. (2020). The impact of monolingual and bilingual subtitles on visual attention, cognitive load, and comprehension. *The Journal of Specialised Translation*, *33*, 70–98.

Lin, P. (2021). Developing an intelligent tool for computer-assisted formulaic language learning from YouTube videos. *ReCALL*. [Advance online publication] https://doi.org/10.1017/s0958344021000252

Lin, P. M. S. (2014). Investigating the validity of internet television as a resource for acquiring L2 formulaic sequences. *System*, *42*(1), 164–176. http://dx.doi.org/10.1016/j.system.2013.11.010

Lin, P. M. S., & Siyanova-Chanturia, A. (2015). Internet television for L2 vocabulary learning. In D. Nunan & J. C. R. Richards (Eds.), *Language learning beyond the classroom* (pp. 149–158). Routledge.

Lindgren, E., & Muñoz, C. (2013). The influence of exposure, parents, and linguistic distance on young european learners' foreign language comprehension. *International Journal of Multilingualism*, *10*(1), 105–129. https://doi.org/10.1080/14790718.2012.679275

Lwo, L., & Chia-Tzu Lin, M. (2012). The effects of captions in teenagers' multimedia L2 learning. *ReCALL*, *24*(2), 188–208. https://doi.org/10.1017/S0958344012000067

Majuddin, E., Siyanova-Chanturia, A., & Boers, F. (2021). Incidental acquisition of multiword expressions through audiovisual materials. *Studies in Second Language Acquisition*, *43*(5), 985–1008. https://doi.org/10.1017/S0272263121000036

Mariotti, C. (2015). A survey on stakeholders' perceptions of subtitles as a means to promote foreign language learning. In Y. Gambier, A. Caimi, & C. Mariotti (Eds.), *Subtitles and language learning. Principles, strategies and practical experiences* (pp. 83–103). Peter Lang.

Markham, P., Peter, L., & McCarthy, T. (2001). The effects of native language vs. target language captions on foreign language students' DVD video comprehension. *Foreign Language Annals*, *34*(5), 439–445. https://doi.org/10.1111/j.1944-9720.2001.tb02083.x

Matthews, J., Cheng, J., & O'Toole, J. M. (2014). Computer-mediated input, output and feedback in the development of L2 word recognition from speech. *ReCALL*, *27*(3), 321–339. http://doi.org/10.1017/S0958344014000421

Mayer, R. E. (2001). *Multimedia learning*. Cambridge University Press.

Mayer, R. E., Fiorella, L., & Stull, A. (2020). Five ways to increase the effectiveness of instructional video. *Educational Technology Research and Development*, *68*, 837–852. https://doi.org/10.1007/s11423-020-09749-6

Mayer, R. E., Lee, H., & Peebles, A. (2014). Multimedia learning in a second language: A cognitive load perspective. *Applied Cognitive Psychology*, *28*(5), 653–660. https://doi.org/10.1002/acp.3050

Mirzaei, M., Meshgi, K., Akita, Y., & Kawahara, T. (2017). Partial and synchronized captioning: A new tool to assist learners in developing second language listening skill. *ReCALL*, *29*(2), 178–199. doi:10.1017/S0958344017000039

Mitterer, H., & McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PLoS One*, *4*(11), 4–8. https://doi.org/10.1371/journal.pone.0007785

Mohd Jelani, N. A., & Boers, F. (2018). Examining incidental vocabulary acquisition from captioned video: Does test modality matter? *ITL - International Journal of Applied Linguistics*, *169*(1), 169–190. https://doi.org/10.1075/itl.00011.jel

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. doi:10.1017/S0272263119000706

Montero Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014a). Effects of captioning on video comprehension and incidental vocabulary learning. *Language Learning & Technology*, *18*(1), 118–141. https://doi.org/10125/44357

Montero Perez, M., Peters, E., & Desmet, P. (2014b). 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. *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., & Rodgers, M. (2019). Video and language learning. *The Language Learning Journal*, *47*(4), 403–406. https://doi.org/10.1080/09571736.2019.1629099

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

Mora, J., & Cerviño-Povedano, E. (2019). The effects of bimodal L2 input on the processing of function words by Spanish EFL learners: An eye-tracking study. In C. Herrero & I. Vanderschelden (Eds.), *Using film and media in the language classroom* (pp. 76–91). Multilingual Matters.

Muñoz, C. (2017). The role of age and proficiency in subtitle reading. An eye-tracking study. *System*, *67*, 77–86. https://doi.org/10.1016/j.system.2017.04.015

Muñoz, C., Cadierno, T., & Casas, I. (2018). Different starting points for English language learning: A comparative study of Danish and Spanish young learners. *Language Learning*, *68*(4), 1076–1109. https://doi.org/10.1111/lang.12309

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

Muñoz, C., Pujadas, G., & Pattemore, A. (2021). Audio-visual input for learning L2 vocabulary and grammatical constructions. *Second Language Research*. [Advance online publication]. https://doi.org/10.1177/02676583211015797

Nakanishi, T. (2015) A meta-analysis of extensive reading research. *TESOL Quarterly*, *49*(1), 6–37. https://doi.org/10.1002/tesq.157

Nation, P. (2007). The four strands. *Innovation in Language Learning and Teaching*, *1*(1), 1–13. https://doi.org/10.2167/illt039.0

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*, 95–106. https://doi.org/10.2307/747835

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

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

Perego, E., Del Missier, F., Porta, M., & Mosconi, M. (2010). The cognitive effectiveness of subtitle processing. *Media Psychology*, *13*(3), 243–272. https://doi.org/10.1080/15213269.2010.502873

Peters, E. (2018). The effect of out-of-class exposure to English language media on learners' vocabulary knowledge. *ITL - International Journal of Applied Linguistics*, *169*(1), 142–168. https://doi.org/10.1075/itl.00010.pet

Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audio-visual input. *TESOL Quarterly*, *53*(4), 1008–1032. https://doi.org/10.1002/tesq.531.

Peters, E. (2020). Factors affecting the learning of single-word items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 125–142). Routledge.

Peters, E., Heynen, E., & Puimège, E. (2016). Learning vocabulary through audio-visual 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., & Muñoz, C. (2020). Introduction to the special issue. Language learning from multimodal input. *Studies in Second Language Acquisition*, *42*(3), 489–497. https://doi.org/10.1017/S0272263120000212

Peters, E., Noreillie, A., Heylen, K., Bulté, B., & Desmet, P. (2019). The impact of instruction and out-of-school exposure to foreign language input on learners' vocabulary knowledge in two languages. *Language Learning*, *69*(3), 747–782. https://doi.org/10.1111/lang.12351z

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

Price, K. (1983). Closed-captioned TV: An untapped resource. *MATSOL Newsletter*, *12*(2), 1–8.

Puimège, E., & Peters, E. (2019a). Learners' English vocabulary knowledge prior to formal instruction: The role of learner-related and word-related variables. *Language Learning*, *69*(4), 943–977. https://doi.org/10.1111/lang.12364

Puimège, E., & Peters, E. (2019b). Learning L2 vocabulary from audio-visual 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., & Peters, E. (2020). Learning formulaic sequences through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition*, *42*(3), 525–549. https://doi.org/10.1017/s027226311900055x

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

Pujadas, G., & Muñoz, C. (2020). Examining adolescent EFL learners' TV viewing comprehension through captions and subtitles. *Studies in Second Language Acquisition*, *42*(3), 1–25. https://doi.org/10.1017/S0272263120000042

Pujolà, J.-T. (2002). CALLing for help: Researching language learning strategies using help facilities in a web-based multimedia program. *ReCALL*, *14*(2), 235–262. https://doi.org/10.1017/S0958344002000423

Quaglio, P. (2010) *Television dialogue. The sitcom friends vs. Natural conversation*. John Benjamins.

Reinders, H., & Benson, P. (2017). Research agenda: Language learning beyond the classroom. *Language Teaching*, *50*(4), 561–578. https://doi.org/10.1017/S0261444817000192

Rodgers, M. (2016). Extensive listening and viewing: The benefits of audiobooks and television. *The European Journal of Applied Linguistics and TEFL*, *5*(2), 43–57.

Rodgers, M. (2018). The images in television programs and the potential for learning unknown words. *ITL - International Journal of Applied Linguistics*, *169*(1), 191–211. https://doi.org/10.1075/itl.00012.rod

Rodgers, M., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly*, *45*(4), 689–717. http://doi.org/10.5054/tq.201

Rodgers, M., & Webb, S. (2017). The effects of captions on EFL learners ' comprehension of English-language television programs. *CALICO Journal*, *34*(1), 20–39. https://doi.org/10.1558/cj.29522

Rodgers, M., & 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

Rose, K. R. (2001). Compliments and compliment responses in film: Implications for pragmatics research and language teaching. *IRAL - International Review of Applied Linguistics in Language Teaching*, *39*(4), 309–326. https://doi.org/10.1515/iral.2001.007

Rost, M. (2002). *Teaching and researching listening*. Pearson Education Limited.

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

Rost, M. (2014). Developing listening fluency in asian EFL settings. In T. Muller, J. Adamson, P. Shigeo Brown, & S. Herder (Eds.), *Exploring EFL fluency in Asia* (pp. 281–296). Palgrave Macmillan.

Ruck, J. (2020). Elementary-level learners' engagement with multimodal resources in two audio-visual genres. *The Language Learning Journal*. [Advance online publication] https://doi.org/10.1080/09571736.2020.1752291

Satar, H. M. (2016). Meaning-making in online language learner interactions via desktop videoconferencing. *ReCALL*, *28*(3), 305–325. https://doi.org/10.1017/S0958344016000100

Sauro, S., & Zourou, K. (2019). What are the digital wilds? *Language Learning & Technology*, *23*(1), 1–7. https://doi.org/10125/44666

Scheffler, P., Jones, C., & Domińska, A. (2021). The Peppa Pig television series as input in pre-primary EFL instruction: A corpus-based study. *International Journal of Applied Linguistics*, *31*(1), 3–17. https://doi.org/10.1111/ijal.12298

Serafini, F. (2012). Expanding the four resources model: Reading visual and multimodal texts. *Pedagogies*, *7*(2), 150–164. https://doi.org/10.1080/1554480X.2012.656347

Sherman, J. (2010). *Using authentic video in the language classroom*. Cambridge University Press.

Sockett, G. (2012). From the cultural hegemony of English to online informal learning: Cluster frequency as an indicator of relevance in authentic documents. *ASp, la revue du GERAS*, *60*, 5–20. https://doi.org/10.4000/asp.2469

Sockett, G., & Kusyk, M. (2015). Online informal learning of English: Frequency effects in the uptake of chunks of language from participation in web-based activities. In T. Cadierno & S. W. Eskildsen (Eds.), *Usage-based perspectives on second language learning* (pp. 153–177). Mouton De Gruyter.

Sokoli, S., & Zabalbeascoa, P. (2019). Audio-visual activities and multimodal resources for foreign language learning. In C. Herrero & I. Vanderschelden (Eds.), *Using film and media in the language classroom* (pp. 170–187). Multilingual Matters.

Suárez, M. M., & Gesa, F. (2019). Learning vocabulary with the support of sustained exposure to captioned video: Do proficiency and aptitude make a difference? *The Language Learning Journal*, *47*(4), 497–517. https://doi.org/10.1080/09571736.2019.1617768

Sweller, J. (2005). The redundancy principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 159–167). Cambridge University Press.

Sydorenko, T. (2010). Modality of input and vocabulary acquisition. *Language Learning & Technology*, *14*(2), 50–73. http://dx.doi.org/10125/44214

Szarkowska, A., Díaz Cintas, J., & Gerber-Morón, O. (2021). Quality is in the eye of the stakeholders: What do professional subtitlers and viewers think about subtitling? *Universal Access in the Information Society*, *20*, 661–675. https://doi.org/10.1007/s10209-020-00739-2

Talaván, N., & Avila-Cabrera (2015). First insights into the combination of dubbing and subtitling as L2 didactic tools. In Y. Gambier, A. Caimi, & C. Mariotti (Eds.), *Subtitles and language learning. Principles, strategies and practical experiences*. Peter Lang.

Taylor, G. (2005). Perceived processing strategies of students watching captioned video. *Foreign Language Annals*, *38*, 422–427. https://doi.org/10.1111/j.1944-9720.2005.tb02228.x

Teng, F. (2019a). Maximizing the potential of captions for primary school ESL students' comprehension of English-language videos. *Computer Assisted Language Learning*, *32*(7), 665–691. https://doi.org/10.1080/09588221.2018.1532912

Teng, F. (2019b). Incidental vocabulary learning for primary school students: The effects of L2 caption type and word exposure frequency. *The Australian Educational Researcher*, *46*(1), 113–136. https://doi.org/10.1007/s13384-018-0279-6

Teng, F. (2019c). The effects of video caption types and advance organizers on incidental L2 collocation learning. *Computers and Education*, *142*, 103655. https://doi.org/10.1016/j.compedu.2019.103655

Teng, F. (2020). Vocabulary learning through videos: Captions, advance-organizer strategy, and their combination. *Computer Assisted Language Learning*, https://doi.org/10.1080/09588221.2020.1720253

Teng, F. (2021). *Language learning through captioned videos: Incidental vocabulary acquisition*. Routledge.

Terantino, J. M. (2011). Emerging technologies YouTube for foreign languages: You have to see this video. *Language Learning & Technology*, *15*(1), 10–16. http://dx.doi.org/10125/44231

The Douglas Fir Group. (2016). A transdisciplinary framework for SLA in a multilingual world. *Modern Language Journal*, *100*(Supplement), 19–47. https://doi.org/10.1111/modl.12301

Tragant, E., Muñoz, C., & Spada, N. (2016). Maximizing young learners' input: An intervention program. *Canadian Modern Language Review*, *72*(2), 234–257. https://doi.org/10.3138/cmlr.2942

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning; A meta-analysis of correlational studies. *Language Learning*, *69*(3), 559–599. https://doi.org/10.1111/lang.12343

Vandergrift, L. (2004). Listening to learn or learning to listen? *Annual Review of Applied Linguistics*, *24*, 3–25. https://doi.org/10.1017/S0267190504000017

Vanderplank, R. (1988). The value of teletext sub-titles in language learning. *English Language Teaching Journal*, *42*(4), 272–281. https://doi.org/10.1093/elt/42.4.272

Vanderplank, R. (1993). A very verbal medium: Language learning through closed captions. *TESOL Journal*, *3*(1), 10–14.

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press

Vanderplank, R. (2010). Déjà vu? A decade of research on language laboratories, television and video in language learning. *Language Teaching*, *43*(1), 1–37. https://doi.org/10.1017/S0261444809990267
Vanderplank, R. (2016a). *Captioned media in foreign language learning and teaching*. Palgrave MacMillan.
Vanderplank, R. (2016b). 'Effects of' and 'effects with' captions: How exactly does watching a TV programme with same-language subtitles make a difference to language learners? *Language Teaching*, *49*(2), 235–250. https://doi.org/10.1017/S0261444813000207
Vanderplank, R. (2019). 'Gist watching can only take you so far': Attitudes, strategies and changes in behaviour in watching films with captions. *The Language Learning Journal*, *47*(4), 407–423. https://doi.org/10.1080/09571736.2019.1610033
Vanderplank, R. (2020). Video and informal language learning. In M. Dressman & R. W. Sadler (Eds.), *The handbook of informal language learning* (pp. 181–201). Wiley-Blackwell.
Van Lommel, S., Laenen, A., & d'Ydewalle, G. (2006). Foreign grammar acquisition while watching subtitled television programs. *British Journal of Educational Psychology*, *76*, 243–258. https://doi.org/10.1348/000709905X38946
Verspoor, M. H., de Bot, K., & van Rein, E. (2011). English as a foreign language. The role of out-of-school language input. In A. De Houwer & A. Wilton (Eds.), *English in Europe today. Sociocultural and educational perspectives* (pp. 147–166). John Benjamins.
Vulchanova, M., Aurstad, L. M. G., Kvitnes, I. E. N., & Eshuis, H. (2015). As naturalistic as it gets: Subtitles in the English classroom in Norway. *Frontiers in Psychology*, *5*, 1–10. https://doi.org/10.3389/fpsyg.2014.01510
Wang, Y. (2012). Learning L2 vocabulary with American TV drama: From the learner's perspective. *English Language Teaching*, *5*(8), 217–225. doi:10.5539/elt.v5n8p217
Wang, Y. (2019). Effects of L1/L2 captioned TV programs on students' vocabulary learning and comprehension. *CALICO Journal*, *36*(3), 204–224. https://doi.org/10.1558/cj.36268
Webb, S. (2010). Pre-learning low-frequency vocabulary in second language television programmes. *Language Teaching Research*, *14*(4), 501–515. http://doi.org/10.1177/1362168810375371
Webb, S. (2015). Extensive viewing: Language learning through watching television. In D. Nunan & J. C. Richards (Eds.), *Language learning beyond the classroom* (pp. 159–168). Routledge.
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*(4), 651–675. https://doi.org/10.1017/S0272263114000606
Webb, S., & Rodgers, M. (2009a). The lexical coverage of movies. *Applied Linguistics*, *30*(3), 407–427. https://doi.org/10.1093/applin/amp010
Webb, S., & Rodgers, M. (2009b). Vocabulary demands of television programs. *Language Learning*, *59*(2), 335–366. https://doi.org/10.1111/j.1467-9922.2009.00509.x
Werner, V. (2020). TV discourse, grammaticality, and language awareness. *TESL-EJ*, *24*(3), 1–23.
Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology*, *14*(1), 65–86. http://dx.doi.org/10125/44203
Winke, P., Sydorenko, T., & Gass, S. (2013). Factors influencing the use of captions by foreign language learners: An Eye-tracking study. *Modern Language Journal*, *97*(1), 254–275. https://doi.org/10.1111/j.1540-4781.2013.01432.x
Wisniewska, N., & Mora, J. C. (2020). Can captioned video benefit second language pronunciation? *Studies in Second Language Acquisition*, *42*(3), 1–26. https://doi.org/10.1017/S0272263120000029
Wong, K. M., & Samudra, P. G. (2019). L2 vocabulary learning from educational media: Extending dual-coding theory to dual-language learners. *Computer Assisted Language Learning*, *34*(8), 1182–1204. https://doi.org/10.1080/09588221.2019.1666150
Yeldham, M. (2018). Viewing L2 captioned videos: What's in it for the listener? *Computer Assisted Language Learning*, *30*(8), 864–883. https://doi.org/10.1080/09588221.2017.1406956

Maribel Montero Perez is an Assistant Professor of second language acquisition and language didactics at Ghent University, Belgium. Her research interests include, amongst others, technology-enhanced listening and how audio-visual input can contribute to vocabulary learning. She has published her research in *Studies in Second Language Acquisition*, *the Modern Language Journal*, *System*, and *Language Learning & Technology*.

Cite this article: 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

https://doi.org/10.1017/S0261444821000501 Published online by Cambridge University Press