THIS IS AN EARLIER VERSION OF THE MANUSCRIPT. FOR THE FINAL VERSION, PLEASE CHECK THE JOURNAL WEBSITE: The published version of this paper should be considered authoritative, and any citations or page references should be taken from it: *ITL International Journal of Applied Linguistics*: https://doi.org/10.1075/itl.21039.pel

Multimodal reading and second language learning

Ana Pellicer-Sánchez

University College London

Abstract

Most of the texts that second language learners engage with include both text (written and/or spoken) and images. The use of images accompanying texts is believed to support reading comprehension and facilitate learning. Despite their widespread use, very little is known about how the presentation of multiple input sources affects the attentional demands and the underlying cognitive processes involved. This paper provides a review of research on multimodal reading, with a focus on attentional processing. It first introduces the relevant theoretical frameworks and empirical evidence provided in support of the use of pictures in reading. It then reviews studies that have looked at the processing of text and pictures in first and second language contexts. Based on this review, main gaps in research and future research directions are identified. The discussion provided in this paper aims at advancing research on multimodal reading in a second language. Achieving a better understanding of the underlying cognitive processes in multimodal reading is crucial to inform pedagogical practices and to develop theoretical accounts of second language multimodal reading.

Keywords: multimodality, second language learning, reading, eye-tracking, cognitive processes, multimedia learning, imagery

Introduction

The ability to read is an essential skill in the development of second language (L2) communicative competence. Reading is a complex cognitive activity that requires the use and integration of different cognitive processes and knowledge bases (Grabe, 2006). Learning to read in an L2 is a long learning process that requires considerable cognitive effort (Grabe, 2006). Thus, a main concern of reading researchers and practitioners has been to find ways to support the L2 reading process, helping learners develop their reading skills and maximising the learning potential of reading.

A common approach to support reading is to combine the presentation of written text with other modes (e.g., images, audio), creating multimodal reading conditions. While the

majority of L2 reading research has traditionally focused on text-only reading conditions, many L2 reading materials include both text and images (e.g., graded readers, stories in EFL textbooks, etc.). In fact, the majority of real-world reading involves the integration of text and pictures (static or dynamic), e.g., newspaper articles, websites, social media posts. The information presented in the written text and in the images is integrated to support reading comprehension. The benefits of this combination of verbal and non-verbal information are usually explained in light of Mayer's (2001, 2009, 2014, 2021) Cognitive Theory of Multimedia Learning, which posits that learners can form better referential connections when verbal and visual materials are presented simultaneously. The verbal information in multimodal reading materials can also be presented auditorily, in what is known as assisted reading or reading-while-listening.

The presentation of different sources of input in multimodal reading materials (i.e., text, images, audio) directly affects the cognitive processes involved and how learners process the text, as the presentation of multiple input sources leads to an added demand on attentional resources. As Grabe (2009) argues, "attentional processing drives many cognitive processes that are critical for reading comprehension" (p. 68). Cognitive processing factors impact the degree of success of reading instruction (Grabe, 2006) and thus, a better understanding of the cognitive processes involved in L2 reading has direct and important implications for L2 reading instruction.

However, despite the importance of understanding attentional processing in multimodal texts, very few studies have investigated learners' processing of information in multimodal reading in the L2 context. The majority of studies examining the processing of multimodal texts involving text and pictures have been conducted in the first language (L1), in the context of science learning (e.g., Johnson & Mayer, 2012; Mason, Pluchino, Tornatora, & Ariasi, 2013; Mason, Tornatora, & Pluchino, 2015), where the main purpose of reading is domain learning. Thus, findings from these studies do not necessarily translate well to other contexts where readers engage with texts for comprehension and for the development of linguistic skills, which are common purposes in L2 reading. Some recent studies have attempted to fill this gap and have examined L2 learners' processing of text and pictures in multimodal texts. These studies suggest that the presence of images attract the attention of both young and adult L2 readers but that they spend most of the time processing the text (e.g., Serrano & Pellicer-Sánchez, 2019; Tragant & Pellicer-Sánchez, 2019). The presence of auditory input also affects processing patterns, allowing L2 readers to better integrate the verbal and non-verbal information in multimodal texts (Pellicer-Sánchez, et al., 2018; Pellicer-Sánchez, et al., 2020; Pellicer-Sánchez, Conklin, Rodgers, & Parente, 2021).

This paper provides a review of relevant research in the area of multimodal L2 reading, with a particular focus on attentional processing. The discussion provided aims at identifying main emerging patterns from the most recent research, as well as main gaps that should be addressed in future investigations to advance research in the area of multimodal L2 reading.

Multimodality in reading

Multimodal texts convey meaning through different modes. Multimodal reading combines the presentation of written text with other modes, usually with static/dynamic pictures and/or auditory input. The definition of modes has been contested and they have been defined differently across fields and schools of thought (Mills & Unsworth, 2017). From a social semiotics perspective, modes include speech, gesture, written language, music, mathematical notation, drawings, photographic images, or moving digital images (Mills & Unsworth, 2017).

The benefits of the use of pictures for literacy development are widely acknowledged. Previous studies have shown that the addition of static pictures to written texts supports reading comprehension and that they have an important role in the development of reading skills (Wright, 2010). Understanding a story text involves the creation of a mental representation of the text, and pictures can facilitate the creation of this mental representation (Boerma, Mol, & Jolles, 2016), by providing an initial framework on which readers can base their mental model (Eitel & Scheiter, 2015). Pictures in reading assist learners in the construction of meaning, helping them keep in mind information about the overall context and about the situations and characters appearing in the text (Wright, 2010). Successful reading comprehension in this type of multimodal texts depends on the ability to integrate the text and images. In fact, this visual-verbal interface has been regarded as a crucial dimension of literacy development (e.g., Bezemer & Kress, 2008; Kress, 2000; Unsworth, 2014).

The benefits of the simultaneous presentation of text and pictures for content learning and literacy development are usually supported by two main cognitive theories, i.e., Paivio's Dual Coding Theory (1986, 2006) and Mayer's Cognitive Theory of Multimedia Learning (2001, 2009, 2014a, 2021). The Dual Coding Theory explains that cognition involves the activation of two systems, i.e., a verbal system that deals with language and a nonverbal system (imagery) that deals with non-linguistic objects and events (Clark & Paivio, 1991). This theory suggests that the verbal and non-verbal input are processed through these two different channels and that the simultaneous activation of the verbal and nonverbal systems fosters learning. Building on the Dual Coding Theory, the Cognitive Theory of Multimedia Learning attempts to explain how people construct meaning from pictures and words. According to this theory, when learners process text and pictures simultaneously, they select relevant words and images, organise them in coherent verbal and pictorial representations, and integrate them with prior knowledge (Mayer, 2014b). Mayer (2009) put forward a list of empirically based multimedia principles that explain successful multimedia learning. Table 1 presents those that are most relevant for the present discussion.

Table 1. *Principles of Multimedia Learning (Mayer, 2021)*

Principle	**Definition**
Multimedia Principle	"People learn better from words and pictures than from words alone" (p.117).
Redundancy Principle	"People do not learn better when printed text is added to graphics and narration" (p.186).

Spatial Contiguity Principle	"People learn better when corresponding words and pictures are presented near rather than far from each other on the page or screen" (p. 207).
Temporal Contiguity Principle	"People learn better when corresponding words and pictures are presented simultaneously rather than successively" (p. 227).
Modality Principle	"People learn more deeply from pictures and spoken words than from pictures and written words" (p. 281).

The main tenet of this theory is the multimedia principle, which posits that people learn better from words and pictures than from words alone. Particularly relevant for the type of materials discussed in this paper (i.e., text + pictures) are the spatial contiguity principle, temporal contiguity principle, the modality principle, and the redundancy principle (see Table 1). Empirical evidence supporting these principles abounds. Research has shown that students who receive both text and pictures perform better in transfer tasks (i.e., tasks in which participants should use the knowledge that they have acquired during a learning phase in a new context) than students who receive words only (*multimedia principle*) (e.g., Mayer & Anderson, 1991; Moreno & Mayer, 1999). In support of the *spatial contiguity principle*, studies have also shown that students who receive integrated presentations (words and pictures next to each other) outperform those who received separated presentations (words and pictures far from each other) (e.g., Johnson & Mayer, 2012; Makransky, Terkildsen, & Mayer, 2019). Similarly, learners who receive simultaneous presentation of narration and animation seem to have an advantage over those who receive the narrative before or after the animation, supporting the *temporal contiguity principle* (e.g., Mayer & Anderson, 1991). Evidence also suggests that students perform better on transfer tests if they receive narration with animation rather than onscreen text with animation, i.e., *modality principle* (e.g., Mayer & Moreno, 1998; Moreno & Mayer, 2002a). Empirical evidence to support the *redundancy principle* is rather mixed. Initial evidence suggested that students who received narration and animation (non-redundant group) had an advantage over students who received narration, animation, and onscreen text (redundant group) (e.g., Moreno & Mayer, 2002b). However, later studies showed that the negative redundancy effects were mitigated when students had control over the pace of the lesson (e.g., Mayer, Howarth, Kaplan, & Hanna, 2018). Research findings to date suggest that the redundancy effect is strongest when learners do not have control over the pace of the presentation (Mayer, 2021).

Crucially, the vast majority of empirical evidence supporting the use of pictures alongside text (multimedia principle) has been conducted in the context of domain learning in the L1 (e.g., science, maths, or geography learning). It is also important to note that the multimedia effect has been mainly found in studies using expository texts, but little evidence has been provided in the case of narrative texts (Boerma, Mol, & Jolles, 2016).

Multimodal reading in second language learning

While multimodal input has been extensively studied in other research disciplines, its effects on L2 learning have only recently attracted the attention of SLA researchers (Peters & Muñoz, 2020). Studies on multimodal L2 input have mainly investigated the effects of captioned/subtitled TV viewing (dynamic images + narration + onscreen subtitles/captions) on comprehension (e.g., Montero Perez, Van Den Noortgate, & Desmet, 2013; Rodgers & Webb, 2017) vocabulary learning (e.g., Puimège, Montero Perez, & Peters, 2021; Rodgers & Webb, 2019; Winke, Gass, & Sydorenko, 2010), and to a lesser extent grammar learning (e.g., Lee & Révész, 2018) (for a critical review of multimodal input in SLA see Montero Perez, 2020). While captioned/subtitled viewing also involves the combination of onscreen text and images, the text plays a supporting role. This section focuses on multimodal reading conditions where the written text has a major role in conveying meaning. Despite their prevalence in everyday input and their potential benefits, far less research has been conducted examining input that involves the presentation of text and static pictures in multimodal reading conditions. They not only seem to be neglected in L2 reading research, but they also seem to be undervalued in the language classroom. Seburn (2017) argued that "despite best intentions, visuals are often an underutilized, cursory or completely absent aspect to reading skills lessons in language learning classrooms, most noticeably at higher levels" (p. 79).

The few empirical studies conducted have shown that the use of static pictures in reading supports reading comprehension (e.g., Elley & Mangubhai, 1983; Omaggio, 1979; Plass, Chun, Mayer, & Leutner, 1998). They also seem to help learners decode words, (e.g., Center, Freeman, Robertson, & Outhred, 1999), which facilitates vocabulary learning (e.g., Bisson, Van Heuven, Conklin, & Tunney, 2015; Plass et al., 1998). Previous studies have also shown that the facilitation effect of images is not only observed in outcome measures, but it is also perceived by the learners (Tang, 1992).

From a theoretical perspective, studies on the effect of multimodal input on L2 learning often interpret their findings in light of Pavio's Dual Coding Theory. According to this theory, the benefits observed from input that includes both auditory/onscreen text and dynamic/static imagery are due to the simultaneous activation of verbal and non-verbal systems. Crucially, findings from L2 multimodal research are also usually interpreted in relation to Mayer's Cognitive Theory of Multimedia Learning. In the absence of a more relevant theoretical account, principles of multimedia learning (see Table 1) are often cited to discuss research findings in the L2 context. However, the applicability of these principles to studies on multimodal L2 learning has been questioned. As argued earlier, these principles were based on empirical evidence gathered from studies on learning from expository texts, which are different from the type of multimodal materials often examined in the L2 context (e.g., videos with onscreen text, narrative texts with pictures). Researchers have also noted the need to distinguish between domain learning and language learning, because of their different purposes and learning goals (Abraham & Farías, 2017; Schnotz & Baadte, 2008). In addition, Mayer's work mainly focused on L1 learning. Researchers have argued that it is unlikely that these principles translate to the L2 learning context without modification. In fact, empirical evidence has demonstrated that some of these principles do not apply to L2 learning contexts. For example, Lee and Mayer (2018) provided evidence against the modality principle. L2 participants in this study watched an instructional video in one of three conditions: video +

narration, video + onscreen text, and video + narration + onscreen text. Results of the comprehension test showed that the video + onscreen text group scored higher than each of the other two groups. In addition, the study provided evidence for a reversed redundancy effect, by showing that participants who watched the narrated video with onscreen text (the redundant group) outperformed students who were only exposed to the narrated video (non-redundant group). Comprehension of auditory input involves the processing of extended samples of L2 speech, automatically and in real time (Buck, 2001). Unlike written input, the ephemeral nature of listening makes it impossible for learners to go back to previous parts of the spoken text. Thus, for L2 learners, redundant onscreen text may help them process the verbal stream by eliminating the need to keep up with the pace of the narration (Mayer, 2021). The onscreen text helps segment individual words from the continuous speech stream. Furthermore, the conditions including onscreen text were rated by learners as less difficult than the condition with narration only. While more research is needed to understand the applicability of the multimedia principles to a range of L2 learning contexts, the boundary conditions that have been demonstrated so far should be considered when interpreting findings from studies on L2 multimodal learning.

Cognitive processing of multimodal texts

The review presented in the previous sections has shown the positive effects of using pictures in L1 and L2 learning, as well as the theoretical accounts of such benefits. However, very little is known about how the presence of multiple input sources affects attentional processing in multimodal reading.

A detailed record of cognitive processing during multimodal reading can be obtained using eye-tracking. Recordings of eye movements reveal the cognitive processes underlying learners' interactions with multimodal texts (Holsanova, 2014). In the last decade, researchers have taken advantage of the affordances of eye-tracking to examine learners' processing of text and pictures in multimodal reading conditions, and a considerable body of research has been accumulated to date. The majority of this research has been conducted in the L1 context. An important aim of these studies has been to explore the relative amount of attention paid to text and images in different multimodal reading conditions. Studies with pre-literate children in shared reading experiences have shown that most of children's attention is paid to the static images, with the text drawing little attention (e.g., Evans & Saint-Aubin, 2005; Justice, Skibbe & Canning, 2005). Attention to the text seems to increase with age and proficiency (e.g., Roy-Charland, Saint-Aubin, & Evans, 2007), and adult, proficient readers seem to allocate most of their attention to the text (e.g., Hannus & Hyönä, 1999; Johnson & Mayer, 2012; Schmidt-Weigand, 2011; Schmidt-Weigand, Kohnert, & Glowalla, 2010). Previous studies have also shown that the use of dynamic visuals attracts more attention than static visuals (e.g., Chen, Hsiao, & She, 2015; Takacs & Bus, 2016).

The majority of eye-tracking studies exploring the processing of text and images in the L1 context have focused on domain/content learning (e.g., Mason, et al., 2013a; Mason, Tornatora, & Pluchino, 2015), with most of them examining the processing of science

materials (Alemdag & Cagiltay, 2018). In this context, studies have shown that learners attend the images only marginally, and that comprehension is mainly driven by the text (Hannus & Hyönä, 1999). The use of abstract illustrations seems to promote more efficient processing of the text than concrete illustrations and leads learners to make a greater effort to integrate verbal and pictorial information (e.g., Mason, et al., 2013a). Importantly, the ability to integrate text and pictures has been shown to support retention and enhance performance (e.g., Mason, Tornatora, & Pluchino, 2015; Mason et al., 2013b).

In the L2 context, very few eye-tracking studies have been conducted to examine learners' cognitive processes during multimodal reading. An important aim of the existing studies has been to explore attention allocation to the text and pictures. Despite claims that images in reading might attract too much attention and distract L2 learners, particularly in the case of young learners (Hill, 2013), recent studies have shown that both young and adult learners allocate most of their attention to the text (e.g., Serrano & Pellicer-Sánchez, 2019; Tragant & Pellicer-Sánchez, 2019; Pellicer-Sánchez, et al., 2020; Pellicer-Sánchez, et al., 2021). While dynamic visuals attract more attention than static images, most of learners' attention is still allocated to the text (e.g., Tragant & Pellicer-Sánchez).

Another significant aim of recent eye-tracking studies in the L2 context has been to examine how the addition of auditory input affects the processing of text and pictures in multimodal reading. In these studies, learners were asked to read an illustrated narrative text in either reading-only (i.e., written verbal input) or reading-while-listening (i.e., written and auditory input) while their eye movements were recorded. The results have consistently shown that, the addition of auditory input to written text and static images leads to increased attention to the pictures. The presence of auditory input allows learners to process the images more often. This pattern seems to be consistent for both young (e.g., Pellicer-Sánchez et al., 2020; Serrano & Pellicer-Sánchez, 2019) and adult readers (Pellicer-Sánchez, et al., 2021). The addition of auditory input also allows for more integrative transitions between text and pictures (Pellicer-Sánchez et al., 2021). Interestingly, the different processing patterns in reading-only and reading-while-listening conditions reported in these studies do not seem to be reflected in comprehension differences.

A further aim of these recent eye-tracking studies has been to examine the relationship between processing patterns and comprehension in multimodal texts. Chang and Choi (2014) were, to the best of my knowledge, the first ones to explore this relationship. They examined the effect that seductive text (i.e., emotionally interesting text segments) and attention-grabbing pictures had on adult L2 learners' reading comprehension and recall of information from the text. The results showed a negative relationship between processing of the seductive segments in the text and comprehension, with more attention related to worse comprehension. Attention to the pictures did not predict recall nor comprehension. A similar negative relationship was reported with young L2 learners by Serrano and Pellicer-Sánchez (2019) and Pellicer-Sánchez et al., (2020), with more time spent processing the text related to lower levels of comprehension. The longer processing time on the text in these studies was therefore interpreted as a reflection of processing difficulties. More recently, Pellicer-Sánchez et al., (2021) reported a different relationship between processing and comprehension for L2

adult learners and L1 adult readers. Their findings suggested that processing time on the text was related to lower comprehension for L1 speakers, but it seemed to be related to better comprehension for L2 readers. Importantly, processing time on static images was positively related to comprehension only for L1 speakers. Overall, existing findings concerning the relationship between processing time and comprehension measures are mixed and they seem to vary by language background and proficiency.

Research Gaps and Future directions

While recent research has positively contributed to our understanding of cognitive processes during multimodal L2 reading, many relevant questions about attentional processing in multimodal reading remain to be answered. In this section, I identify some of the directions that research on multimodal L2 reading should take to further our understanding of how L2 readers process multimodal texts and their effect on performance measures. This evidence is crucial to inform theories of L2 multimodal reading.

Purposes of reading

The first notable gap is that the processing studies conducted in the L2 context have focused on reading for general comprehension. However, L2 readers engage with texts for different purposes (reading for general understanding, reading to learn content, scanning, skimming, etc.), and the cognitive processes involved are determined by the specific purpose. In academic contexts, for example, skimming to get the general gist of the text or scanning to locate specific information are important reading strategies. Reading can also serve multiple purposes simultaneously, as in the case of reading for domain and linguistic learning in Content and Language Integrated Learning (CLIL) contexts. Grabe (2006) explained that "As we read for different purposes, we shift how we use our cognitive processes and knowledge resources" (p. 281). It seems logic to assume that the allocational processing of texts and pictures and the level of integration will depend on the purpose of reading. However, the existing L2 eye-tracking studies have all examined reading for general comprehension. The potential effect that different purposes of reading may have on attentional processes in L2 multimodal reading is yet to be empirically examined.

Relationship between attentional allocation and learning

The available eye tracking studies exploring the relationship between attention allocation and L2 outcome measures have focused on comprehension. Findings from these studies seem to suggest that processing differences found in several multimodal texts (i.e., written text + picture; written text + pictures + auditory text) are not reflected in comprehension differences (e.g., Serrano & Pellicer-Sánchez, 2019; Pellicer-Sánchez, et al., 2020; Pellicer-Sánchez, et al., 2021). However, it is unknown whether differences in attention allocation could be reflected in differences in learning gains. L2 reading studies involving the presentation of written text only (e.g., Godfroid, Boers, & Housen, 2013; Pellicer-Sánchez, 2016) and those exploring viewing (e.g., Montero Perez, Peters, & Desmet, 2015; Puimège, Montero Perez, & Peters, 2021) have suggested a relationship between processing patterns and vocabulary

learning gains. Whether a similar relationship is found in multimodal reading is yet to be examined.

Image-Text relationship

In multimodal texts, images and text coexist to convey meaning. The image is considered an independent message that is connected to the verbal text but not dependent on it (Kress & van Leeuwen, 2021). The level of complexity of the meaning relations that text and image hold can vary. Images can simply illustrate what the text relates, or they can expand the information related in the text and even contradict it (Guijarro & Sanz, 2009). Several frameworks have been provided to explain the various relations between text and image. For example, Guijarro and Sanz (2009), building on research on systemic functional linguistics, identified four main types of intersemiotic relations between text and images: *symmetrical interaction* (i.e., words and image convey the same story, repeating information), *ideational complementarity* (i.e., words and images convey different but complementary information), *counterpointing interplay* (i.e., words and images convey alternative information), and *contradictory interaction* (i.e., words and images contradict each other). Similarly, building on social semiotics systems, Unsworth (2006) identified three main possibilities for the construction of ideational meaning in multimodal texts: *ideational concurrence* (i.e., the image and text have similar participant-process-phenomenon configuration); *ideational complementarity* (i.e., what is represented in the images and text is different but complementary); and *connection* (i.e., quoting and reporting of speech or thoughts, and portraying casual, temporal, and spatial relations).

Most of the L2 eye-tracking studies reviewed in this paper explored the processing of narrative texts in which pictures supported the information presented in the text, having mainly a symmetrical and concurrent relationship. The integration of text and pictures and the cognitive processes involved in this integration are bound to depend on the relationship between text and image. Therefore, these frameworks should be used in future studies to explore the processing of images and texts that hold different intersemiotic relations, providing a more comprehensive understanding of the integration of input sources in multimodal texts.

Conclusion

L2 learners often engage with multimodal texts that include both texts and pictures. The integration of text and pictures in multimodal conditions has been shown to support reading comprehension and enhance learning. The review provided in this paper has demonstrated that, despite the widespread use of imagery in L2 texts and the importance of understanding the attentional processes involved, very few studies have examined the processing of text and pictures in multimodal L2 reading. The few available studies have shown that both young and adult L2 learners allocate most of their attention to the text, and that the addition of auditory input (in combination with written input and images) leads to longer processing time on the images and more transitions between text and image, allowing for their better integration. These processing differences do not seem to have a detrimental effect on comprehension, supporting the role of multimodality for L2 learning. Initial evidence has also been provided for the relationship between processing time on the text/images and comprehension, with

different patterns reported according to proficiency. While recent evidence has contributed to our understanding of the cognitive processes involved in multimodal reading, important questions remain to be addressed. This paper has argued that, in order to reach an accurate understanding of attentional processing in multimodal reading, future research needs to investigate the relation between cognitive processes and learning gains, as well as the processing of multimodal texts with different purposes of reading and with texts and images that hold various intersemiotic relations. Notably, gathering this empirical evidence is crucial to be able to build a more comprehensive model of L2 multimodal and multimedia learning.

References

Abraham, P., & Farías, M. (2017). Reading with eyes wide open: Reflections on the impact of multimodal texts on second language reading. *Íkala, Revista de Lenguaje y Cultura, 22*(1), 57–70. doi:10.17533/udea.ikala.v22n01a04

Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education, 125,* 413–428. doi:10.1016/j.compedu.2018.06.023

Bezemer, J., & Kress, G. (2008). Writing in multimodal texts: A social semiotic account of designs for learning. *Written Communication, 25*(2), 166–195. doi:10.1177/0741088307313177

Bisson, M-J., W. Van Heuven, K. Conklin & R. Tunney. 2015. The role of verbal and pictorial information in multimodal incidental acquisition of foreign language vocabulary. *Quarterly Journal of Experimental Psychology 68*, 1306–1326. doi:10.1080/17470218.2014.979211

Boerma, I. E., Mol, S. E., & Jolles, J. (2016). Reading Pictures for Story Comprehension Requires Mental Imagery Skills. *Frontiers in Psychology, 7* (1630), 1-10. doi:10.3389/fpsyg.2016.01630

Buck, G. (2001). *Assessing Listening*. New York, NY: Cambridge University Press.

Center, Y., Freeman, L., Robertson, G., & Outhred, L. (1999). The effect of visual imagery training on the reading and listening comprehension of low listening comprehenders in year 2. *Journal of Research in Reading 22*, 241–256. doi.org/10.1111/1467-9817.00088

Chang, Y., & Choi, S. (2014). Effects of seductive details evidenced by gaze duration. *Neurobiology of Learning and Memory 109*, 131–138. doi:10.1016/j.nlm.2014.01.005

Chen, S. C., Hsiao, M. S., & She, H. C. (2015). The effects of static versus dynamic 3D representations on 10th grade students' atomic orbital mental model construction: Evidence from eye movement behaviors. *Computers in Human Behavior, 53*, 169–180. doi: 10.1016/j.chb.2015.07.003

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review, 3*, 149-210. doi:10.1007/BF01320076

Ediger, A. M. (2006). Developing strategic L2 readers…by reading for authentic purposes. In Usó-Juan, E. & Martínez-Flor, A. (Eds.), *Current trends in the development and teaching of the four language skills* (pp. 303-328). New York: Mouton de Gruyter.

Eitel, A., & Scheiter,K. (2015). Picture or text first? Explaining sequence effects when learning with pictures and text. *Educational Psychology Review, 27,* 153–180. doi:10.1007/s10648-014-9264-4

Elley, W. B. & F. Mangubhai (1983). The impact of reading on second language learning. *Reading Research Quarterly 19*(1), 53–67. doi:10.2307/747337

Evans, M. A. & J. Saint-Aubin (2005). What children are looking at during shared storybook reading: Evidence from eye movement monitoring. *Psychological Science 16,* 913–920. doi:10.1111/j.1467-9280.2005.01636.x

Grabe, W. (2006). Areas of research that influence L2 reading instruction. In Usó-Juan, E. & Martínez-Flor, A. (Eds.), *Current trends in the development and teaching of the four language skills* (pp. 279-301). New York: Mouton de Gruyter.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice.* Cambridge: Cambridge University Press.

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye-tracking. *Studies in Second Language Acquisition, 35*(3), 483-517. doi:10.1017/S0272263113000119

Guijarro A.J.M., & Sanz M.J.P. (2009) On Interaction of Image and Verbal Text in a Picture Book. A Multimodal and Systemic Functional Study. In Ventola E., & Guijarro A.J.M. (Eds.), *The world told and the world shown* (pp. 107-123). London: Palgrave Macmillan.

Hannus, M., & Hyönä, J. (1999). Utilization of Illustrations during Learning of Science Textbook Passages among Low- and High-Ability Children. *Contemporary Educational Psychology, 24*(2), 95–123. doi:10.1006/ceps.1998.0987

Hill, D. (2013). Graded readers. *ELT Journal 67*(1). 85–125. doi:10.1093/elt/ccs067

Holsanova, J. (2014). Reception of multimodality: Applying eye tracking methodology in multimodal research. In C. Jewitt (Ed.), *The Routledge Handbook of Multimodal Analysis* (pp. 287–298). New York: Routledge.

Johnson, C. I., & Mayer, R. E. (2012). An eye movement analysis of the spatial contiguity effect in multimedia learning. *Journal of Experimental Psychology: Applied, 18,* 178–191. doi:10.1037/a0026923

Justice, L. M., Skibbe, L., Canning, A., & Lankford, C. (2005). Pre-schoolers, print and storybooks: An observational study using eye movement analysis. *Journal of Research in Reading 28,* 229–243. doi:10.1111/j.1467-9817.2005.00267.x

Kress, G. (2000). *Multimodality: A Social Semiotic Approach to Communication*. London: Routledge.

Kress, G. & van Leeuwen, T. (2021). *Reading images: The grammar of visual design*. NY: Routledge.

Lee, H., & Mayer, R. (2018). Fostering learning from instructional video in a second language. *Applied Cognitive Psychology, 32*, 648-654. doi:10.1002/ACP.3436

Lee, M., & Révész, A. (2018). Promoting grammatical development through textually enhanced captions: An eye-tracking study. *The Modern Language Journal, 102*, 557–577. https://doi.org/10.1111

Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Role of subjective and objective measures of cognitive processing during learning in explaining the spatial contiguity effect. *Learning and Instruction, 61*, 23-34. doi: 10.1016/j.learninstruc.2018.12.001

Mason, L., Pluchino, P., Tornatora, M. C., & Ariasi, N. (2013a). An eye-tracking study of learning science text with concrete and abstract illustrations. *Journal of Experimental Education 81,* 356–384. doi:10.1080/00220973.2012.727885

Mason, L., Tornatora, M. C., & Pluchino, P. (2013b). Do fourth graders integrate text and picture in processing and learning from an illustrated science text? Evidence from eye-movement patterns. *Computers and Education 60,* 95–109. doi:10.1016/j.compedu.2012.07.011

Mason, L., Tornatora, M. C., & Pluchino, P. (2015). Integrative processing of verbal and graphical information during re-reading predicts learning from illustrated text: An eye movement study. *Reading and Writing, 28,* 851–872. doi:10.1007/s11145-015-9552-5.

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

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge: Cambridge University Press.

Mayer, R. E. (2014a). Introduction to multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed.) (pp. 1–24). Cambridge: Cambridge University Press.

Mayer, R. (2014b). Cognitive Theory of Multimedia Learning. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 43–71). Cambridge: Cambridge University Press.

Mayer, R. E. (2021). *Multimedia learning* (3rd ed.). Cambridge: Cambridge University Press.

Mayer, R. E., & Anderson, R. B. (1991). Animations need narrations: An experimental test of a dual-coding hypothesis. *Journal of Educational Psychology, 83*(4), 484–490. doi:10.1037/0022-0663.83.4.484

Mayer, R. E., Howarth, J. T., Kaplan, M., & Hanna, S. (2018). Applying the segmenting principle to online geography slideshow lessons. *Educational Technology Research and Development, 66*(3), 563–577. doi:10.1007/s11423-017-9554-x

Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology, 90*(2), 312–320. doi:10.1037/0022-0663.90.2.312

Mills, K., & Unsworth, L. (2017). *Multimodal Literacy*. Oxford Research Encyclopedia of Education. Retrieved 22 Jul. 2021, from: https://oxfordre.com/education/view/10.1093/acrefore/9780190264093.001.0001/acrefore-9780190264093-e-232.

Montero Perez, M. (2020). Multimodal input in SLA research. *Studies in Second Language Acquisition, 42*(3), 653-663. doi:10.1017/S0272263120000145

Montero Perez, M., Van Den Noortgate, W., & Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System, 41*, 720–739. doi:10.1016/j.system.2013.07.013

Montero Perez M., Peters, E., & Desmet, P. (2015) Enhancing vocabulary learning through captioned video: an eye-tracking study. *The Modern Language Journal 99*(2), 308–28. doi:10.1111/modl.12215

Moreno, R., & Mayer, R. E. (1999) Multimedia-supported metaphors for meaning making in mathematics. *Cognition and Instruction, 17*(3), 215-248. doi: 10.1207/S1532690XCI1703_1

Moreno, R., & Mayer, R. E. (2002a). Learning science in virtual reality multimedia environments: Role of methods and media. *Journal of Educational Psychology, 94*(3), 598–610. doi:10.1037/0022-0663.94.3.598

Moreno, R., & Mayer, R. E. (2002b). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology, 94*(1), 156–163. doi:10.1037/0022-0663.94.1.156

Omaggio, A. C. (1979). Pictures and second language comprehension: Do they help? *Foreign Language Annals 12,* 107-116. doi:10.1111/j.1944-9720.1979.tb00153.x

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

Paivio, A. (2006). *Mind and its evolution: A dual coding approach*. Lawrence Erlbaum.

Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition from and while reading: An eye-tracking study. *Studies in Second Language Acquisition, 38*, 97–130. doi: 10.1017/S0272263115000224.

Pellicer-Sánchez, A., Conklin, K., Rodgers, M., & Parente, F. (2021). The Effect of Auditory Input on Multimodal Reading Comprehension: An Examination of Adult Readers' Eye Movements. Early view: https://doi.org/10.1111/modl.12743

Pellicer-Sánchez, A., Tragant, E., Conklin, K., Rodgers, M., Llanes, A., & Serrano, R. (2018). L2 reading and reading-while-listening in multimodal learning conditions: An eye-tracking study. *ELT Research Papers, 18*(01), 1-28. London: British Council.

Pellicer-Sánchez, A., Tragant, E., Conklin, K., Rodgers, M., Serrano, R., & Llanes, A. (2020). Young learners' processing of multimodal input and its impact on reading comprehension: An eye-tracking study. *Studies in Second Language Acquisition, 42*(3), 577-598. doi:10.1017/S0272263120000091

Peters, E., & Muñoz, C. (2020). Introduction to the special issue: Language learning from multimodal input. *Studies in Second Language Acquisition, 42*(3), 489-497. doi:10.1017/S0272263120000212

Plass, J.L., Chun, D.M., Mayer, R.E., & Leutner, D. (1998). Supporting visual and verbal learning preferences in a second language multimedia learning environment. *Journal of Educational Psychology, 90*, 25–36. doi: 10.1037/0022-0663.90.1.25

Puimège, E., Montero Perez, M., & Peters, E. (2021). Promoting L2 acquisition of multiword units through textually enhanced audiovisual input: An eye-tracking study. *Second Language Research*. Online First. doi:10.1177/02676583211049741

Rodgers, M. P. H., & Webb, S. (2017). The effects of captions on EFL learners' comprehension of English-language television programs. *Calico Journal, 34*, 20–38. doi:10.1558/cj.29522

Rodgers, M. P. H. & Webb, S. (2019). Incidental vocabulary learning through viewing television. *ITL-International Journal of Applied Linguistics, 171*(2), 191-220. doi:10.1075/itl.18034.rod

Roy-Charland, A., Saint-Aubin, J., & Evans, M. A. (2007). Eye movements in shared book reading with children from kindergarten to grade 4. *Reading and Writing 20*, 909–931. doi:10.1007/s11145-007-9059-9

Schmidt-Weigand, F. (2011). Does animation amplify the modality effect–or is there any modality effect at all? *Zeitschrift für Padagogische Psychologie, 25*(4), 245–256. doi: 10.1024/1010-0652/a000048

Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010). A closer look at split visual attention in system-and self-paced instruction in multimedia learning. *Learning and Instruction*, *20*(2), 100–110. doi: 10.1016/j.learninstruc.2009.02.011

Schnotz, W., & Baadte, C. (2008). Domain learning versus language learning with multimedia. In M. Farías & K. Obilinovic (Eds.), *Aprendizaje multimodal/Multimodal learning* (pp. 21-49). Santiago de Chile: Publifahu USACH.

Seburn, T (2017). Learner-sourced visuals for deeper text engagement and conceptual comprehension. In K. Donaghy & D. Xerri (Eds.), *The image in English language teaching* (pp. 79-88). Malta: Gutenberg Press.

Serrano, R., & Pellicer-Sánchez, A. (2019). Young L2 learners' online processing of information in a graded reader during reading-only and reading-while-listening conditions: A study of eye movements. *Applied Linguistics Review*. Advance online publication. doi:10.1515/applirev-2018-0102

Takacs, Z. K., & Bus, A. G. (2016). Benefits of motion in animated storybooks for Children's visual attention and story comprehension. An eye-tracking study. *Frontiers in Psychology, 7,* 1–12. doi:10.3389/fpsyg.2016.01591

Tang, G. (1992). The effect of graphic representation of knowledge structures on ESL reading comprehension. *Studies in Second Language Acquisition 14*, 177–195.

Tragant, E., & Pellicer-Sánchez, A. (2019). Young learners' engagement with multimodal exposure: An eyetracking study. *System, 80*, 212–223. doi:10.1016/j.system.2018.12.002.

Unsworth, L. (2006). Towards a metalanguage for multiliteracies education: Describing the meaning—making resources of language-image interaction. *English Teaching, 5*(1), 55–76.

Unsworth, L. (2014). Multimodal reading comprehension: Curriculum expectations and large-scale literacy testing practices. *Pedagogies: An International Journal, 9*(1), 26–44. doi:10.1080/1554480X.2014.878968

Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology, 14,* 65–86. doi:10125/44214.

Wright, A. (2010). *Pictures for language learning*. Cambridge: Cambridge University Press.