Intersemiotic Cohesion between Texts and Images in the Dialogues of the *Stretch 1* English Textbook

Mareeyadar Teeda-aksornsakul[1,*]
Parada Dechapratumwan[2]
Saran Chantalay[3]

*(**Received:** June 19, 2025; **Revised:** January 12, 2026; **Accepted:** January 14, 2026)*

Abstract

This study examines intersemiotic cohesion between verbal texts and visual images in the dialogues of *Stretch 1*, an English commercial textbook. Liu and O'Halloran's (2009) intersemiotic cohesive devices: Reference, Theme-Rheme Development, Given-New Organization, and Parallel Structure were used in the research to analyze 12 dialogues from all units of the book to assess how these elements interact to convey meaning. Results show partial coherence and inconsistencies in intersemiotic cohesion. Reference and Theme-Rheme cohesion were most frequent, while Parallel Structure was least consistent. This underscores the need for stronger text-image integration to support upper-beginner learners.

Keywords: Cohesive devices; ELT textbook; intersemiotic cohesion; multimodality; text-image relations

[1,*] Lecturer Dr., Western Languages Division, Faculty of Liberal Arts, Maejo University (Corresponding author)
E-mail: mareedar.maew@gmail.com
[2] Assistant Professor Dr., Western Languages Division, Faculty of Liberal Arts, Maejo University
E-mail: koiparada@gmail.com
[3] Lecturer Dr., Western Languages Division, Faculty of Liberal Arts, Maejo University
E-mail: Saran_Chantalay@hotmail.com

Introduction

Communication has increasingly relied on multiple channels, and this shift has extended beyond the media industry to influence educational contexts as well, particularly in textbooks using various modes, such as visuals and audio, to reinforce content and facilitate learning. However, there are not many studies examining how verbal and visual modes interact to construct meaning in commercial textbooks used to teach EFL learners at the upper-beginner level. To address this gap, this study analyzes the *Stretch 1* textbook, a commercial textbook from Oxford University Press used by many Thai universities, as a representative example of a multimodal textbook that is explicitly integrating images with dialogues and listening tasks, which is a suitable corpus for exploring how visual grammar contributes to meaning-making and language learning at the beginner level for EFL students, by using intersemiotic cohesive devices proposed by Liu and O'Halloran (2009). To this end, the findings aim to deepen the understanding of multimodal design in English language teaching materials, together with providing pedagogical insights for designing more effective visual-verbal integration in ELT textbooks.

Multimodality is an interdisciplinary field that examines how various semiotic modes: linguistic, visual, and others collaborate to communicate effectively (Kress & Van Leeuwen, 2001). Within this framework, intersemiotic cohesion, the semantic linkage between different modes, becomes critical to ensure that meanings are clear and accessible.

Research Objectives

This study examines the interaction between verbal text and image in the dialogues of the *Stretch 1* textbook to determine how cohesion is established. Therefore, the study addresses two primary objectives: (1) to identify how visual and verbal elements interact to create cohesion, and (2) to analyze the degree of cohesion between images and language in the dialogues of the speaking sections. The findings can illuminate the effectiveness of text-image relations in educational materials and guide future textbook development.

Literature Review

From Linguistic Cohesion to Intersemiotic Cohesion

Cohesion is the lexical and grammatical connections that unify elements of a text into a unified whole (Halliday & Hasan, 1976). This concept has expanded beyond linguistic discourse to encompass multimodal communication, wherein coherence between verbal and non-verbal elements is essential. Van Leeuwen (1991) extended the notion of cohesion beyond linguistic boundaries and emphasized its importance in texts integrating visual and textual modes. Scholars like Royce (1998) explore these relations in scientific articles and advertisements and introduce intersemiotic complementarity, the way different modes, such as visual and verbal modes, work together to construct meaning, developed from Halliday and Hasan's (1976) cohesion in the English language and visual grammar of Kress and Van Leeuwen (1996). According to Royce (1998), an image and its accompanying text cohere through repetition, synonymy, antonymy, meronymy, hyponymy, and collocation. For example, an image of fire coheres with the text 'fire' through repetition as it has an identical meaning, while it coheres with the text 'passion' through synonymy when it is used metaphorically to convey desire, for example, 'She spoke with fire in her voice.'

Liu and O'Halloran (2009) extended Halliday and Hasan's framework to define intersemiotic cohesion as the semantic interplay among different modes in multimodal texts. They affirm that intersemiotic texture is a semantic connection across various modes generated by intersemiotic cohesive devices in multimodal communication, which creates coherence to produce meaning. They also adopt Royce's (1998) intersemiotic complementarity to identify

lexical semantic relations, Hasan's (1985) Parallelism, a device tracing parallel meanings between texts to identify shared structural patterns across modes, including implication sequence (O'Halloran, 2005), looking at logical relations between modes, and Jones's (2008) Correspondence to identify semantic alignment between modes. Additionally, they adapt Reference, Theme-Rheme Development, and Given-New Organization from Kress and Van Leeuwen's (2006), and introduce Parallel Structure to find how text and visual or spatial resources jointly organize information as Intersemiotic Cohesive Devices (see details in 3.1 Theoretical Framework).

Intersemiotic Cohesive Devices

According to Liu and O'Halloran (2009), four cohesive devices: Reference, Theme-Rheme Development, Given-New Organization, and Parallel Structure can be used to analyze intersemiotic cohesion. Reference is used to identify participants across modes. For Theme-Rheme, it assesses salience and structure, while Given-New evaluates how information is placed, and Parallel Structure examines transitivity alignment. These tools provide a robust framework for dissecting text-image relationships in multimodal texts that can be applied to teaching and learning materials, including coursebooks. They also allow educators to assess the quality of multimodal cohesion in order to make the cohesion teachable and accessible to teachers and learners.

However, there are insufficient studies that have applied Liu and O'Halloran's cohesive devices, particularly for analyzing English language teaching materials. Existing applications remain limited to some aspects of O'Halloran's cohesive devices, for example, Parallel Structure, Reference, Implication Sequences, and Correspondence (See Alyousef, 2021; Attar, 2014; Khalid et al., 2017; Royce, 2008; Teedaaksornsakul, 2024). A recent study by Teedaaksornsakul (2024) demonstrates how intersemiotic cohesion between images and texts plays an important role in ESP materials for EFL learners. The study found that the intersemiotically cohesive visuals serve as scaffold tools to improve EFL students' understanding of technical vocabulary and procedural steps and increase learners' motivation to engage in learning activities. In addition, when visual images were semantically linked to texts, they connected abstract technical texts with tangible representations, making it more comprehensible to infer the meanings, and also reducing the cognitive load. While their findings offer valuable insights, the analysis focuses on Parallel Structure, which leaves the other cohesive devices underexplored. Consequently, there is a need for research to apply other intersemiotic cohesive devices to evaluate multimodal cohesion in teaching and learning materials, particularly those designed for beginner-level English language learners.

Multimodal Cohesion in Educational Practice

In educational contexts, cohesion in the design of teaching material plays a central role in ensuring that diverse semiotic resources, such as texts, typography, and images, are balanced to convey messages (Lim & Nguyen, 2022). Due to this, the interplay of various modes used to enhance multimodal literacy has become the main focus in pedagogical research (e.g., Jewitt, 2008; Lim & Toh, 2020; Teedaaksornsakul & Bowen, 2024). For instance, Jewitt (2008) points out that it is vital for teachers and students to have the ability to explain and interpret choices of semiotic resources for designing teaching and learning materials, how they interact to convey meaning, and multimodal cohesion in meaning-making. Likewise, Lim (2018) affirms the importance of cohesive multimodal texts together with teaching multimodal literacy. Additionally, Lim and Tan-Chia (2023) propose a metalanguage used for analyzing and describing multimodal cohesion to help teachers explicitly teach how elements like images and language align in structure and meaning. However, their frameworks need a full tool for

micro-level analysis to analyze how cohesion is realized between modes, which enables researchers to examine intersemiotic relations in learning materials.

Multimodal Textbooks and the Research Gap

Many researchers have explored how multimodal textbooks shape learning outcomes and support meaning-making across different subject areas, such as Guo (2004), Chen (2010), and Salbego et al. (2015). They have demonstrated that semantic coordination between textual and visual elements enhances comprehension, particularly in English as a Foreign Language (EFL) instruction. Their findings suggest that well integration between verbal language and visuals facilitates students' understanding of abstract concepts, difficult vocabulary, and develops grammatical awareness.

Recent studies have emphasized analyzing and developing multimodal textbooks using multimodal theory (e.g., Teedaaksornsakul, 2024; Park et al., 2025; Yanuar et al., 2021). For example, the study of Yanuar et al. (2021) indicates that the interplay between clothing and contexts, gestures and facial expressions, color symbolism, shadows and realism, and spatial composition and texts strengthens multimodal literacy. However, some symbolism is ambiguous, and mismatches between facial expressions and textual descriptions are confusing.

Research Methodology

Data Collection

Data comprised all 12 dialogues with images from *Stretch 1*'s speaking sections from all 12 units of the textbook. They were selected for their relevance to CEFR at the A2 level. More importantly, this textbook was selected as it is a widely distributed series used by many Thai universities for fundamental English courses. The textbook has diverse images representing everyday situations and social interactions, including dialogues aimed at teaching vocabulary and conversational skills for novice learners; they are ideal for cohesion analysis.

Data Analysis

Liu and O'Halloran's (2009) cohesive devices were systematically applied. These included: Reference, distinguishing between specific (e.g., named participants) and generic (e.g., 'a chef') links; Theme-Rheme Development analyzed through framing via size, color, and position, and salience analyzed through connectedness between an organizational structure of an image and text. The Given-New Organization was evaluated through information value, based on layout, for example, top or left bottom as given information, while bottom or right as new information. Parallel Structure was examined by comparing verbal and visual transitivity processes.

The study is limited to the dialogues in *Stretch 1*, analyzed through Liu and O'Halloran's (2009) cohesive devices, focusing on intersemiotic relationships between verbal and visual modes.

Theoretical Framework

Transitivity System

Halliday's (1994) transitivity system classifies linguistic processes, or how language represents human experiences. Material, mental, relational, behavioral, verbal, and existential processes are six process types that are associated with specific participants and circumstances that enable a structured analysis of how meaning is constructed in discourse. Material processes refer to actions and events, mental processes to perception and cognition, relational processes to state of being or identification, behavior processes to physiological and

psychological behavior, verbal processes to acts of communication, and existential processes to expressions of existence.

This framework is grounded in Halliday's (1994) Systemic Functional Linguistics (SFL). Drawing on Halliday's transitivity system, Kress and Van Leeuwen (2006) extended it to visual processes. Their model identifies analogous visual processes that represent action, reaction, classification, and symbolic meanings in images. For example, actional processes, a subcategory of narrative processes (see the Data Analysis Section for details and explanation) in visuals realized by a vector that indicates movement, are conceptually similar to material processes in language. Likewise, reactional processes are identified through gaze or interaction between represented participants that resemble mental processes. Table 1 illustrates the parallels between Halliday's (1994) linguistic transitivity system and visual transitivity structures of Kress and Van Leeuwen (2006).

Table 1

The Parallel Structures between Halliday's (1994) Transitivity System and Kress and Van Leeuwen's (2006) Visual Transitivity System

Halliday's (1994) Transitivity System		Kress and Van Leeuwen's Visual Transitivity System
Process type	material	narrative processes (action & reaction processes)
	relational • attributive ('being' & 'having') • identifying	conceptual processes • classificational process • analytical process • symbolic attributive process
	Existential	Symbolic suggestive process
circumstances		circumstances

Note. Adapted from Motta-Roth & Nascimento (2009)

Visual Analysis

Visual Grammar categorizes images into Narrative and Conceptual (state of being and having) processes, using vectors and participant roles to interpret meaning. There are five types of narrative processes: action, reactional, verbal, mental, conversion, and circumstance, under narrative processes that represent a participant's action or a movement (Kress & Van Leeuwen, 2006). For conceptual processes, they are identified as classificational (participants' taxonomies), analytical (carriers and possessive attributes), and symbolic (carrier as a symbolic attribute implying connotative meaning) processes.

Cohesion Analysis

Liu and O'Halloran's (2009) framework integrates ideational (logical/

experiential) and textual (Reference, Theme-Rheme, Given-New, Parallel Structure) metafunctions to assess multimodal cohesion. Intersemiotic Reference, a cohesive device, establishes connections between elements across modes. Linguistic elements, such as pronouns, are an example of Reference that points to visual participants or vice versa. For example, a caption stating, 'He is waiting for a train at a train station,' refers to a man standing in a train station represented in an image. For Theme-Rheme Development, it is a mechanism that organizes information. The theme is the starting point of the message, while the Rheme is the new information expanding on the Theme. In terms of the Given-New structure, the Given information is prior, while the New information adds information. Given

and New align with Theme-Rheme. Given aligns with Theme, while New aligns with Rheme as expansion. According to Kress and Van Leeuwen (2006), the image's placement within the layout determines its information value. Any visual elements placed on the left composition are interpreted as Given, while the ones on the right are New information. Similarly, the top of an image is associated with ideas or idealized information (Ideal), while the bottom represents concrete details (Real). When visual (Given) information is followed by verbal (New) information, a cohesive flow is established. This study adopts this model to examine how the image-text compositions distribute meaning through salience and spatial orientation. For the Parallel Structure, it is identified from the parallelism of transitivity processes in text and image. For example, a caption in a dialogue states, 'The teacher is writing the word 'patience' on the whiteboard,' has a Parallel Structure to that of the image representing an action process of the teacher (actor) writing (process) on the whiteboard (circumstance) where the word 'patience' (goal) appears near the teacher's hand holding a marker.

Definitions of Cohesive Levels

In this study, the strength of Reference was categorized into Clear and Weak based on the degree of participant identification, and the degree of intersemiotic cohesion in Parallel Structure was categorized into four levels: Strong, Moderate, Partial, and Weak, adapted from Liu and O'Halloran's (2009) typology of cohesive realization. These categories were further refined and developed as a descriptive coding scale for the dataset of this study. The label 'Clear' Reference is used to describe instances where participants or entities explicitly represented in the image match the linguistic referents in the dialogue. For the label 'Weak' Reference, it is used to refer to instances where one or more linguistic participants are visually absent or only inferentially represented in either the image or the dialogue. Regarding the degree of intersemiotic cohesion in Parallel Structure, Table 2 summarizes the definition of each cohesive level.

Table 2

Intersemiotic Cohesive Levels of Parallel Structure

Cohesive Level	Definition
Strong cohesion	Occurs when verbal and visual modes show precise alignment in transitivity structure, participant roles, and emotional expression.
Moderate cohesion	Occurs when visual and verbal modes share general thematic alignment but diverge in structural or participant-level correspondence.
Partial cohesion	Occurs when images align thematically or contextually with the dialogue but fail to fully represent participant roles or verbal transitivity structures.
Weak cohesion	Occurs when dialogues exhibit significant mismatches between verbal and visual modes, particularly in transitivity structure, participant interaction, and representational meaning.

Results and Discussion

Overview of Findings

Analysis of 12 multimodal dialogues in the *Stretch 1* textbook revealed partial intersemiotic cohesion between visual and linguistic modes. The strongest patterns of intersemiotic cohesion were identified from Reference and Theme–Rheme Development, while the most frequent weaknesses were found in Parallel Structure, particularly in terms of transitivity misalignment between verbal and visual processes.

This section is organized according to the two research objectives. First, the Interaction between Visual and Verbal Elements Section addresses the first objective, focusing on how visual and verbal elements interact to create intersemiotic cohesion. Then, the Degree of Cohesion between Images and Text Section corresponds to the second objective, examining the degree of cohesion between images and language in the speaking dialogues. The analysis follows the four cohesive devices outlined by Liu and O'Halloran (2009), summarized in Table 3 below.

Table 3

Cohesion Analysis Summary

Figure	Reference	Theme-Rheme	Given-New	Parallel Structure
Figure 1	Partial (Vanessa & Sam identified, art generic)	Theme: Image; Rheme: Dialogue (salient image, disconnected framing)	Image: Given; Dialogue: New (adds major info)	Weak
Figure 2	Partial (Tim specific, Lauren absent)	Theme: Image; Rheme: Dialogue (Tim salient, dialogue adds)	Image: Given (Tim driving); Dialogue: New (context)	Weak
Figure 3	Partial (Eric identified, Jen ambiguous)	Theme: Eric's image; Rheme: Narration	Given (outing); Dialogue: New (weekend story)	Moderate
Figure 4	Partial (Tina & Mia identified; male participant not linguistically represented)	Theme: Image (Tina's pointing action providing direction); Rheme: Dialogue	Given (direction); Dialogue: New (details)	Moderate
Figure 5	Partial (Place-specific, speakers absent)	Theme: Bridge; Rheme: Trip narration	Image: Given (places); Dialogue: New (activities)	Partial
Figure 6	Partial (David and kids specific, Sasha missing)	Theme: Image; Rheme: Dinner plan	Image: Given (kids, dinner); Dialogue: New (context)	Moderate
Figure 7	(Jen specific, others absent)	Theme: Jen's image; Rheme: Stress explanation	Image: Given (stress scene); Dialogue: New (workload)	Strong
Figure 8	(Dylan specific, Clara missing, food mismatched)	Theme: Dylan's image; Rheme: Job detail	Image: Given (chef); Dialogue: New (emotion, job)	Moderate

Figure 9	Partial (Carmen assumed, Robert missing)	Theme: Backpackers; Rheme: Trip detail	Image: Given (backpacking); Dialogue: New (destination)	Moderate
Figure 10	Clear (Amy and Laura both identifiable)	Theme: Laura's image; Rheme: Dialogue	Image: Given (Laura sick); Dialogue: New (cold)	Strong
Figure 11	Partial (Jazz club specific, speakers absent)	Theme: Musicians; Rheme: Invitation	Image: Given (musicians); Dialogue: New (plans)	Partial
Figure 12	Partial (Gandhi specific, speakers absent)	Theme: Gandhi's image; Rheme: Traits in dialogue	Image: Given (Gandhi); Dialogue: New (attributes)	Weak

Interaction between Visual and Verbal Elements

The findings found the three cohesive devices: Reference, Theme-Rheme Development, and Given-New Organization, which show how meaning is co-constructed through semiotic interaction. Reference cohesion was categorized from 'Clear' to 'Weak' as defined in the Methodology Section.

Reference

Participant identification was largely successful across figures, especially when names and roles in the dialogue corresponded clearly to visual representations. For example, in Figure 1, participant identification is cohesive: 'Vanessa' and 'Sam' are introduced as specific/presuming participants through proper nouns in the verbal text. These roles are reinforced by the foregrounded images of a man and a woman, making them also specific/presuming in the visual mode. There is no explicit link between the painting image and the verbal text. However, it offers a contextual link to the word 'art' in the dialogue and serves as specific/presuming representations that reinforce the generic/presenting concept of 'art history.' Overall, the alignment between verbal and visual references supports intersemiotic cohesion.

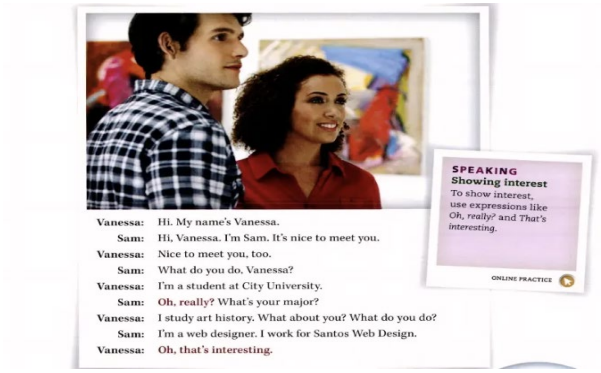

Figure 1: The Analyzed Image and Dialogue from Unit 1 (Stempleski, 2019)

In cases where only one speaker is visually represented, intersemiotic cohesion is partially weakened. In Figure 2, Tim is introduced as a specific/presuming participant both verbally and visually. His presence is emphasized through framing, attire, and salience in the image. However, Lauren, although a named speaker in the dialogue, is visually absent. According to Liu and O'Halloran (2009), visual absence of a specific participant weakens intersemiotic cohesion, particularly when that speaker plays a continuous, agentive role in the discourse (as Lauren does here). The presence of a Bluetooth headset offers a weak inferential

link to her role, but its cohesiveness is weak. Therefore, the alignment between text and image is partial, with a clear representation of Tim but only an implied reference to Lauren.

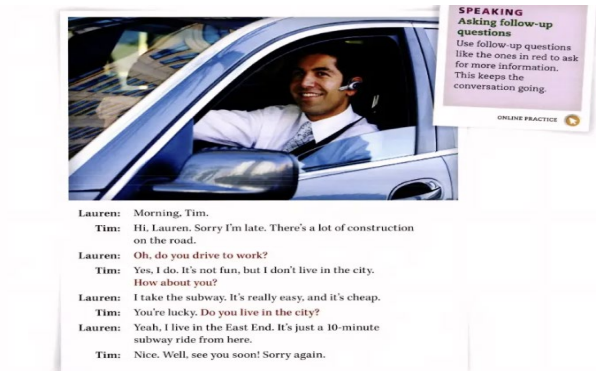

Figure 2: The Analyzed Image and Dialogue from Unit 2 (Stempleski, 2019)

Or when participant identity was ambiguous, such as Jen in Figure 3 or Mia in Figure 4, where multiple similar figures led to uncertainty. In Figure 3, intersemiotic cohesion is only partially realized in terms of participant reference. The male participant, 'Eric,' is introduced in the dialogue as a specific/presuming referent and is visually represented as the only male in the image, positioned centrally and shown interacting with food alongside three women. His salience and participation strongly support his identification across both modes (Halliday & Hasan, 1976; Liu & O'Halloran, 2009). In contrast, 'Jen' is not visually identifiable. The three women in the image could be interpreted as Eric's sister and her friends, as referenced in his verbal narrative, rather than as co-participants in the depicted conversation. Thus, while 'Jen' functions as a specific/presuming participant in the verbal mode, her visual referent is absent or, at best, generic/presenting. According to Liu and O'Halloran (2009), such asymmetry disrupts reference alignment and weakens intersemiotic cohesion. The image, in this case, serves more as a representational Rheme, illustrating Eric's recounted weekend activity, than as a literal depiction of the immediate verbal interaction. Consequently, the reference cohesion is strong for Eric but notably weak for Jen, limiting the overall cohesion of the multimodal ensemble.

Figure 3: The Analyzed Image and Dialogue from Unit 5 (Stempleski, 2019)

In Figure 4, intersemiotic reference is partially cohesive. The female participant 'Tina' is introduced in the dialogue and is visually identifiable as the woman in the middle, pointing at a map next to the male participant. Her interaction with the object (map) supports her function as a specific/presuming participant in both the verbal and visual modes (Halliday & Hasan, 1976; Liu & O'Halloran, 2009). Also, the woman on the left, holding the map, can be identified as 'Mia', who is asking for the Regal Movie Theater in the dialogue. However, the male participant is not mentioned in the dialogue, limiting the full referential alignment.

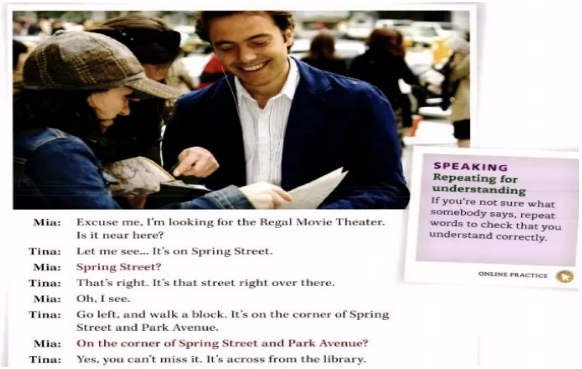

Figure 4: The Analyzed Image and Dialogue from Unit 6 (Stempleski, 2019)

In Figure 5, there is the visual absence of the speakers 'Ron' and 'Keiko,' who are introduced in the dialogue as specific/presuming participants. 'San Francisco,' the 'Golden Gate Bridge,' 'cable cars,' and 'neighborhoods' are mentioned in the dialogue. These references are shown in the image as a specific/presuming participant due to their direct link to the dialogue. While no human participants from the exchange are depicted, referential cohesion is nonetheless maintained through environmental and object alignment. As Liu and O'Halloran (2009) argue, such semiotic reinforcement through place imagery can uphold intersemiotic texture even in the absence of interpersonal referents. Thus, Figure 5 exemplifies how spatial context and iconic signs can function as cohesive bridges between linguistic and visual modes.

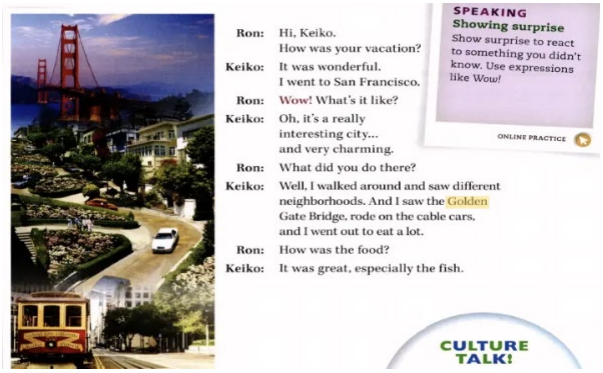

Figure 5: The Analyzed Image and Dialogue from Unit 10 (Stempleski, 2019)

Theme–Rheme Development

Most of the figures serve as the Theme due to their salience in position, size, and color, while the dialogues function as the Rheme. For example, in Figure 1, the composition and the image's salience establish it as the Theme, while the dialogue positioned under the image is the Rheme. This finding is in line with Liu and O'Halloran's (2009) model of intersemiotic cohesion. As seen, Sam's plaid shirt and Vanessa's red blouse contrast with a blurred background. In contrast, the dialogue, placed at the bottom within a white frame, receives less visual emphasis and thus functions as the Rheme. However, despite the logical visual-verbal flow from Theme to Rheme, framing makes them disconnected and reduces the cohesiveness of the multimodal ensemble. According to Kress and Van Leeuwen's (2006) principle, framing weakens unity across semiotic modes. Consequently, although Theme–Rheme Development is structurally present, the intersemiotic cohesion is partial due to compositional disjunction.

Figure 6: The Analyzed Image and Dialogue from Unit 8 (Stempleski, 2019)

Moreover, we also found that some images and their dialogues are not well connected, as their vertical arrangement and framing disconnect between Theme and Rheme. For example, as seen in Figure 6, the image functions as the Theme because of its size, central position, and colors of the children's clothing and vibrant colors of the displayed food, which draw the viewer's attention. According to Liu and O'Halloran (2009), Theme is the mode that initiates the semiotic experience through prominence, whereas Rheme elaborates or expands upon this visual cue. However, while the image effectively anchors the setting and participants, a degree of framing disconnect is introduced by the sharp edge and white background of the dialogue box placed beneath the image. This visual boundary spatially separates the two semiotic layers, diminishing the cohesion between Theme and Rheme (Kress & Van Leeuwen, 2006). As a result, although the dialogue clarifies details such as ingredients and preferences, it appears secondary to the visual, both spatially and perceptually. The image initiates the thematic focus, children enjoying food, while the verbal mode provides elaborative content in the form of a shopping conversation between adults. Therefore, the Theme–Rheme relationship is semiotically established but somewhat disrupted by compositional segmentation.

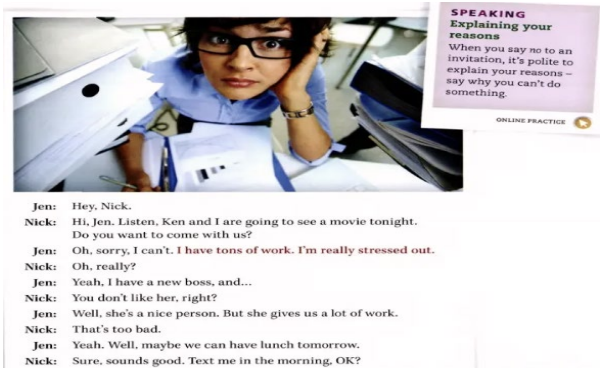

Figure 7: The Analyzed Image and Dialogue from Unit 4 (Stempleski, 2019)

In Figure 7, the image serves as the Theme under its salience and compositional dominance, while the dialogue functions as the Rheme, offering elaboration and contextualization of the visual narrative. The central participant, assumed to be Jen, is visually emphasized by her large-scale placement at the center and top of the frame, directly facing the viewer with a distressed facial expression. Her widened eyes, hand clutching her hair, and furrowed brow act as indexical signs of stress, immediately drawing attention to her emotional state. Surrounding her are chaotic visual cues: loosely stacked folders and papers, along with her upward gaze and slouched posture, all reinforcing the verbal content, 'I have tons of work. I'm really stressed out.' These elements position the image as the viewer's initial point of attention; therefore, the visual Theme, while the verbal Rheme, placed beneath, provides detailed confirmation and extension of the implied narrative. Although the other participant,

Nick, is not visually represented, the image still supports the communicative goal of the dialogue, which is to foreground and explain Jen's emotional and professional strain. The white background of the dialogue, isolated from the image via color contrast and slight spatial distancing, creates some framing separation (Kress & Van Leeuwen, 2006), but this does not critically undermine cohesion. Instead, the image's clear encoding of the vocabulary target ('stressed out') serves as the central thematic focus, with the dialogue elaborating its interpersonal and contextual meanings. Thus, Figure 7 demonstrates effective Theme–Rheme Development, where the visual mode potentially guides the viewer's focus and the verbal mode extends the interpretive meaning of the image within the multimodal composition. Nonetheless, many images created strong thematic grounding that was expanded upon in the dialogue, for example, Dylan's job satisfaction in Figure 8, or backpacking in Figure 9.

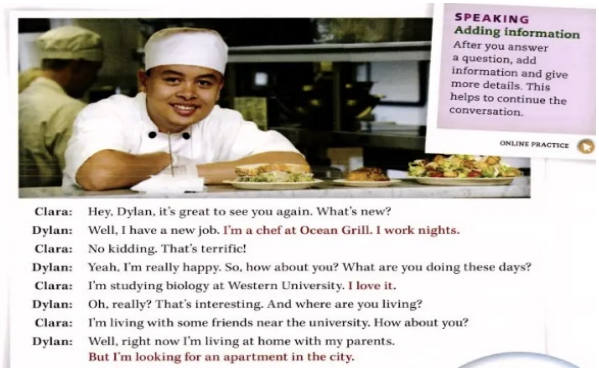

Figure 8: The Analyzed Image and Dialogue from Unit 3 (Stempleski, 2019)

Similarly, in Figure 8, the image is the Theme prominently featuring Dylan, a chef, and his professional setting. The image of Dylan wearing a bright white chef's uniform in the left position is salient because of the size and color contrast against the darker kitchen background. For the dialogue below the image, it functions as the Rheme as it expands on the Theme by providing information about Dylan's new job, his satisfaction with his work, and his living situation. Despite this alignment, a visual framing separation exists between the image and the text, created by a dark edge and compositional contrast between the color-saturated image and the white dialogue box (Kress & Van Leeuwen, 2006). However, this does not significantly impair semiotic cohesion because the dialogue directly elaborates on the image's communicative content.

Likewise, the image of Figure 9 functions as the Theme due to its salience and directional composition. The saturated colors of the three women's T-shirts and backpacks contrast with the vibrant natural background, showing a greenery mountainscape that draws immediate viewer attention. The positions of the three women create a diagonal line leading the eye through the image of the first woman on the left to the last (on the right) to reinforce this thematic function (Kress & Van Leeuwen, 2006), introducing the idea of travel and outdoor adventure. For the dialogue functioning as the verbal Rheme creates a cohesive semiotic unit as it contributes narrative and specific details that Carmen, one of the interlocutors, 'is going to go backpacking' and plans to 'go kayaking' during her vacation in Canada, while Robert is 'going to Thailand.' However, the image possibly introduces Carmen's story as it matches with her vacation activity described in the dialogue.

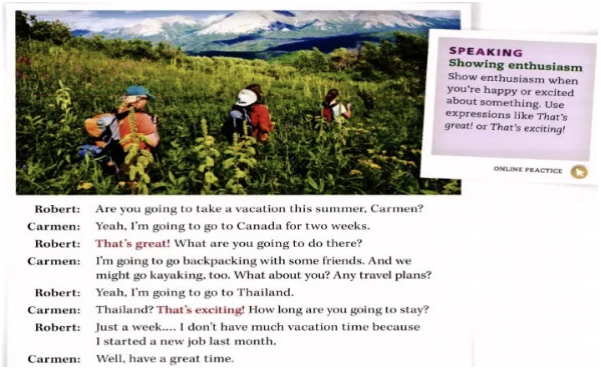

Figure 9: The Analyzed Image and Dialogue from Unit 12 (Stempleski, 2019)

Given–New Organization

For the top-bottom layout, most images function as Given information to provide initial context, for instance, participants, actions, or locations. The dialogues typically function as New information to introduce new circumstances or elaborate roles of the participants. For example, Figure 1 (see the image in the Reference Section) provides the Given information representing the setting information and the participants, Vanessa and Sam. As shown, the paintings in the background indicate that the participants are in an art-related place, which loosely links to the verbal mention of 'art history' and 'design' in the dialogue. However, the image does not directly represent the participants' professional identities as described in the dialogue. Thus, the details about their personal and academic details shown in the dialogue function as the New information. The image only introduces a general thematic setting and does not fully visualize the specific content about their occupations as described in the dialogue.

Also, Figure 2 functions similarly to Figure 1, as the Given information is conveyed through the visual salience of the participant. Tim's central position in a car, his hand on the steering wheel, and the Bluetooth headset on his ear, including his facial expression, possibly suggest that he is engaging in a phone conversation. Then, the dialogue introduces the New information that elaborates on his reason for his lateness ('a lot of construction on the road') and the commuting experiences of Tim and the other speaker. This image aligns with Liu and O'Halloran's (2009) view that images often function as experiential Given information, while texts provide the new information.

Some horizontally arranged figures are also found, and their Given–New distribution followed a left–right orientation. For example, in Figure 5, the image serves as the Given information, offering visual context through a left-right composition featuring three San Francisco landmarks: the Golden Gate Bridge at the top, a winding road with a white car in the center, and a cable car at the bottom. The composite layout draws the viewer's attention primarily to the bright white car in the center due to its contrast with surrounding color tones, while the iconic red of the bridge and the saturated yellow of the cable car further increase the visual complexity. According to Kress and Van Leeuwen (2006), such compositional hierarchy guides the viewer's visual path, making the figure itself the initial point of engagement. However, although the image introduces thematic cues about the location, it does not fully anchor the variety of details presented in the dialogue. While the bridge and cable car are specifically mentioned by Keiko, other verbal references, such as walking through neighborhoods and eating fish, lack visual counterparts, indicating only partial referential cohesion. In these cases, visual salience guided viewer expectations even before dialogue reading. Where New information referred to absent or abstract elements, the cohesion weakened.

Degree of Cohesion between Images and Text

The analysis of Parallel Structure focused on the degree of cohesion between images and text, specifically the extent to which verbal and visual transitivity processes aligned or not aligned. Different degrees of alignment were identified as follows:

Strong Alignment

From the analysis, strong intersemiotic cohesion occurs when verbal and visual modes show precise alignment in transitivity structure, participant roles, and emotional expression. For example, in Figure 10, the parallel structure between verbal and visual transitivity processes is notably strong, marking it as one of the most cohesive examples in the dataset. The dialogue contains Relational attributive processes such as 'You look really bad' and 'I have a bad cold,' where Laura functions as the carrier, and the expressions 'really bad' and 'a bad cold' serve as Attributes (Halliday, 1994). These verbal clauses are realized visually through multiple indexical signs. We can see that Laura's hand raised to her forehead, while her fingers pressed against her brow and nose bridge, and her closed eyes are commonly associated with illness, physical distress, or fatigue. According to Kress and Van Leeuwen (2006), these visual elements represent Analytical or Symbolic attributive processes. The participant's body language and gestures function as visual Attributes to reinforce the verbal description of her condition. Also, the direction of the second participant's gaze (Amy) and her face direction create an eye-line vector directing to Laura's face, functioning as a reactional process. Additionally, the image showing workspace, papers, and computer aligns with the verbal reference to 'so much work,' as described in the dialogue. Thus, the parallelism between the verbal clauses and the image shows a high degree of intersemiotic cohesion.

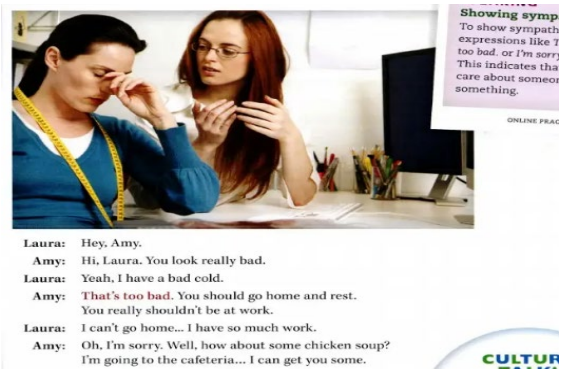

Figure 10: The Analyzed Image and Dialogue from Unit 9 (Stempleski, 2019)

Moderate Alignment

Moderate intersemiotic cohesion occurs when visual and verbal modes share general thematic alignment but diverge in structural or participant-level correspondence. For example, Figure 3 (Eric and Jen) also demonstrates moderate alignment between visual and verbal transitivity structures. In the verbal dialogue, Eric recounts a past event: 'On Saturday, I went out to eat with my sister and her friends,' which is a material process, 'I' as the actor and 'went out to eat' as the process, with 'my sister and her friends' as circumstantial participants. The image represents the event described; however, it does not show the interaction between Eric and Jen. We can assume that the man at the center is Eric from the position and salience. Also, his interaction, including posture and gestures, can indicate that he is eating while the women surrounding him can be his sister and his sister's friends. From a visual transitivity perspective, the image shows action processes. The arm movements and the forks pointing to food items identify the participants as actors and the food as goals. These elements align with Eric's narrative about eating out. Additionally, facial

expressions of smiling while eating support the affective tone of the verbal text ‘we had a good time.’ However, we cannot identify who Jen is from these three women due to the lack of clear referential cues. This limits participant reference cohesion and weakens full alignment. According to Liu and O’Halloran (2009), such partial cohesion is common when visual and verbal modes share thematic but not structural or participant-level correspondence. Therefore, Figure 3 offers a moderate degree of intersemiotic cohesion through consistent thematic alignment and visible material processes, as it lacks full parallelism.

Similarly, Figure 4 (Mia and Tina) is also another example of moderate transitivity alignment. The image partially parallels the verbal content in terms of action and spatial context. The first part of the dialogue is described using the material process, ‘I’m looking for the Regal Movie Theater,’ indicating an action of walking and searching for the place. The speaker is the actor, and ‘the Regal Movie Theater’ is the goal. However, the image only depicts the actor without the goal. The image of the woman pointing at the map while the others are looking at it represents a reactional process. The map functions as the phenomenon while the participants are the sensers. This disjunction between visual reactional and verbal material processes indicates partial intersemiotic cohesion. The other details described, for example, the directions ‘Go left, and walk a block…’ are not visualized in the image but are indirectly referenced through the embodied interaction with the map. In this case, the participants’ engagement with the map enacts the cognitive and spatial orientation functions embedded in the verbal interaction. Therefore, while the specific goal of the material process is not represented, the overarching communicative activity, seeking and giving directions, is visually reinforced, justifying its classification as moderately cohesive.

Partial Alignment

Unlike weak intersemiotic cohesion, partial intersemiotic cohesion occurs when images align thematically or contextually with the dialogue but fail to fully represent participant roles or verbal transitivity structures, as seen in Figures 5 and 11. Figures 5 (Ron and Keiko) and 11 (Ian and Jane) evidence that the absence of speakers and actions resulted in transitivity mismatches. For example, in Figure 5, intersemiotic alignment is only partially achieved, with inconsistencies between the transitivity structures of the verbal and visual modes. The dialogue contains multiple mental and material processes that narrate Keiko’s travel experience. The clause ‘I walked around and saw different neighborhoods’ comprises two Mental processes, where ‘I’ is the senser, and the phenomenon includes ‘different neighborhoods’ and the ‘Golden Gate Bridge.’ Visually, the panoramic image composition includes depictions of the Golden Gate Bridge at the top and a stylized street view at the center, with buildings and cars. While these visual elements offer environmental context relevant to the verbal references, they lack specificity; no individual participant is identifiable as Keiko. This limits interpersonal alignment and reduces cohesion at the participant level. The statement ‘I saw the Golden Gate Bridge’ is more successfully supported by the visual content. The mental process (‘I saw’) aligns with the specific/presuming visual representation of the Golden Gate Bridge, which acts as the phenomenon. The bridge’s salience and iconic design reinforce its status as the referential anchor for the dialogue. However, the alignment weakens when considering the clause ‘I rode on the cable car,’ which is a material process (with ‘I’ as the actor and ‘cable car’ as the goal). Although a cable car is pictured at the bottom of the image, no human participant appears to be engaging with it, breaking the actional congruence. Similarly, the verbal clause ‘I went out to eat’ also reflects a material process, with the participant as the actor and the eating event implied as the process. However, this action is not represented in the image. There are no participants, images of food, or restaurants to represent material action processes through vectors realized by body movement or interaction with objects to support semantic alignment.

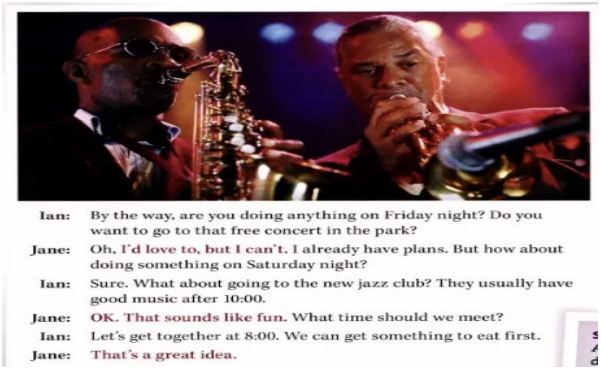

Figure 11: The Analyzed Image and Dialogue from Unit 11 (Stempleski, 2019)

In the same vein, the visual-verbal alignment of Figure 11 is only with a minor connection that is established through lexical overlap. The clause, for example, 'Do you want to go to that free concert in the park?' represents a material process as 'want' and 'go' describe an intended action, 'you' functions as the actor, and 'the park' (a location) is the Circumstance. In contrast, the image depicts two musicians performing jazz. Although the image of jazz can presumably indicate the location of a jazz club, the main actors (Ian and Jane) are absent. Also, there are no contextual visual markers to suggest the other specific locations mentioned in the dialogue, such as 'park' or 'concert venue.' Therefore, no participant-level coherence is established. Instead, there is only a lexical match between the term 'jazz club' and the jazz musicians in the image.

From a transitivity perspective, the image suggests either a behavioral or an action process, showing musicians are playing instruments. However, these actions are not related to the main content of the conversation, which centers on planning and social intention rather than performance. As such, the visual process does not parallel any verbal clause involving participants, goals, or activities. In contrast to Figures with strong or moderate alignment, Figure 11 relies on thematic association (i.e., jazz music) rather than structural cohesion. According to Kress and Van Leeuwen's (2006) framework, true intersemiotic cohesion is achieved not only through topical links but through transitivity correspondence, a convergence of visual and linguistic processes.

Weak Alignment

Among all categories, weak intersemiotic cohesion was the most frequently observed pattern, with five dialogues exhibiting significant mismatches between verbal and visual modes, particularly in transitivity structure, participant interaction, and representational meaning. For example, Figure 1 exemplifies weak intersemiotic cohesion in terms of transitivity alignment. The dialogue includes both relational (e.g., 'I'm Vanessa,' 'I'm Sam') and mental processes (e.g., 'Nice to meet you'), in which the participants function as carriers and sensers, and expressions of emotion or identity function as Attributes or phenomena (Halliday, 1994). These interpersonal and identifying exchanges are foundational to greeting interactions. However, the visual realization of the scene disrupts this cohesion. The two participants, presumably Vanessa and Sam, are positioned in the foreground of the image but are both looking in the same direction, creating a reactional process in which both serve as reactors, and the phenomenon lies outside the visual frame of the painting. Thus, there is no vector connecting the two participants to indicate their interaction as described in the dialogue. The absence of mutual gaze between the two participants contradicts the affective expression 'Nice to meet you.' When participant interaction is verbally emphasized but visually diffused, intersemiotic cohesion is compromised. Thus, the figure reflects low intersemiotic cohesion,

offering an illustrative case of how visual structure can dilute or even negate the communicative function of verbal processes when transitivity systems diverge.

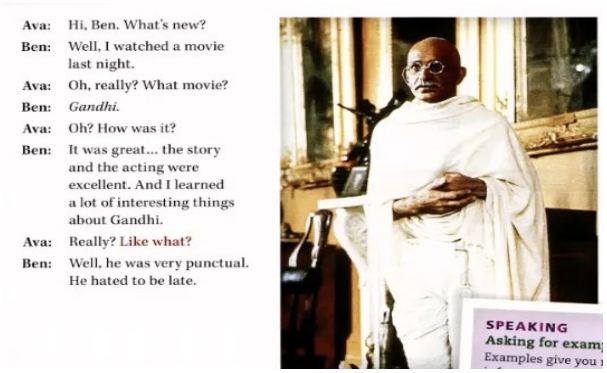

Figure 12: The Analyzed Image and Dialogue from Unit 7 (Stempleski, 2019)

Another example of weak intersemiotic cohesion is seen in Figure 12, as the alignment between verbal and visual transitivity structures is weak. The dialogue focuses on describing Ben's experience of watching a film about Gandhi, which makes Gandhi function as the goal. The dialogue focuses on relational and mental processes, describing Ben's affective stance and Gandhi's personality attribution. For example, the statement 'He was very punctual' shows a relational attributive process. 'He' (Gandhi) is the carrier and 'punctual' is the Attribute (Halliday, 1994). The sentence 'He hated to be late' is a mental process. 'He' is the senser, 'hate' is the mental process, and 'to be late' is the phenomenon.

However, the visual representation of Gandhi in the image does not reflect these verbal processes. The image captures Gandhi in a static pose, with no gesture, action, or emotional expression corresponding to punctuality or emotional attitude. Visually, the figure of Gandhi could be interpreted as either part of a narrative structure (if we assume he is looking at something, implying a reactional process) or as part of a conceptual structure (i.e., symbolic or analytical) as described by Kress and Van Leeuwen (2006). Additionally, his posture and neutral facial expression, including the museum-like background, suggest a symbolic representation. Gandhi can be a symbolic sign of historical significance rather than being in a specific activity. While the dialogue represents Gandhi's personal and psychological states, the visual mode does not show these qualities.

Conclusion

This study examined the intersemiotic cohesion of multimodal dialogues in *Stretch 1* using Liu and O'Halloran's (2009) intersemiotic cohesive devices: Reference, Theme–Rheme, and Parallel Structure. The analysis reveals that Reference and Theme-Rheme Development are the most consistently realized cohesive devices. The visual elements, such as participants, objects, and settings, were found to correspond to their linguistic counterparts. Salient image features, such as the placements of the characters and facial expressions, were found to align with the topics of their dialogues or referential content. However, when analyzing the transitivity alignment from examining Parallel Structure, many visual process types do not match the process types of the verbal texts. Many images do not depict some of the main participants or the interlocutors of the dialogues, or misrepresent the experiential meaning conveyed by the clauses. Although the textbook was created for English language beginners, salient images and referential alignment are insufficient to support multimodal comprehension needed for facilitating English language learning, especially for EFL learners. Strong intersemiotic cohesion can help EFL novice learners comprehend meanings co-constructed between different modes (e.g., Alyousef, 2021; Teedaaksornsakul & Bowen, 2024).

Implications and Recommendations for Future Research

The findings advocate for more intersemiotic cohesion between visuals and verbal texts designed for novice EFL learners. Curriculum and textbook developers should consider alignment between visuals and texts more closely. Further research might explore interpretive strategies and cognitive load of learners when engaging with multimodal materials and expand this analytical approach to include other modes or interactive digital teaching materials in order to illuminate additional aspects of multimodal learning in EFL contexts.

References

Attar, M. M. (2014). *Intersemiotic cohesion analysis of multimodal elements in Iranian English textbooks* [Doctoral dissertation, University of Malaya]. Universiti Malaya Repository. https:// studentsrepo.um.edu.my/5414/

Alyousef, H. S. (2021). Text cohesion in English scientific texts written by Saudi undergraduate dentistry students: A multimodal discourse analysis of textual and logical relations in oral biology texts. *SAGE Open, 11*(3), 1–13. https://doi.org/10.117/21582440211032194

Chen, Y. (2010). The semiotic construal of attitudinal curriculum goals: Evidence from EFL textbooks in China. *Linguistics and Education, 21*(1), 60–74.

Guo, L. (2004). Multimodality in a biology textbook. In K. L. O'Halloran (Ed.), *Multimodal discourse analysis* (pp. 196–219). Continuum.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Halliday, M. A. K. (1994). *Functional Grammar*. Edward Arnold.

Jewitt, C. (2008). Multimodal and literacy in school classrooms. *Review of Research in Education, 32*(1), 241–267. https://doi.org/10.3102/0091732X07310586

Jones, J. (2008). *Multiliteracies for academic purposes: A metafunctional exploration of intersemiosis and multimodality in university textbooks and computer-based learning resources in science* [Doctoral dissertation, University of Sydney]. Sydney eScholarship Repository. https://ses.library.usyd.edu.au/handle/2123/2259

Khalid, S., Malik, M. A., & Butt, M. (2017). An Intersemiotic analysis of verbal and visual cohesion in primary- level textbooks. *Pakistan Languages and Humanity Review, 1*(1), 38–47. http://doi.org/10.47205/plhr.2017(1-I)1.4

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

Kress, G., & Van Leeuwen, T. (1996). *Reading images: The grammar of visual design*. Routledge.

Kress, G., & Van Leeuwen, T. (2006). *Reading images: The grammar of visual design* (2nd ed.). Routledge.

Lim, F. V. (2018). Developing a systemic functional approach to teach multimodal literacy. *Functional Linguistics, 5*, Article 13. https://doi.org/10.1186/s40554-018-0066-8

Lim, F. V., & Toh, W. (2020). How to teach digital reading? *Journal of Information Literacy, 14*(2), 24–43. https://doi.org/10.11645/14.2.2701

Lim, F. V., & Nguyen, T. T. H. (2022). Design-based research approach for teacher learning: A case study from Singapore, *ELT Journal, 76*(4), 452–464. https://doi.org/10.1093/elt/ccab039

Lim, F. V., & Tan-Chia, L. (2023). *Designing learning for multimodal literacy*: *Teaching viewing and representing*. Routledge.

Liu, Y., & O'Halloran, K. L. (2009). Intersemiotic texture: Analyzing cohesive devices between language and images. *Social Semiotics, 19*(4), 367–388.

Motta-Roth, D., & Nascimento, F. S. (2009). Transitivity in visual grammar: Concepts and applications. *Linguagem & Ensino, 12*(2), 319–349. https://doi.org/10.15210/rle.v12i2.15713

O'Halloran, K. L. (2005). *Mathematical discourse: Language, symbolism and visual images*. Continuum.

Salbego, N., Heberle, V. M., & da Silva Balen, M. G. S. (2015). A visual analysis of English textbooks: Multimodal scaffolded learning. *Calidoscópio,13*(1), 5. https://doi.org/10.4013/cld.2015.131.01

Park, J., Lee, Y.J., Koh, A., & Tan, G. (2025). The complementary roles of visual and written representations within figures in science textbooks. *Research in Science Education, 25*(1). https://doi.org/10.1007/s11165-025-10246-0

Royce, T. (1998). Synergy on the page: Exploring intersemiotic complementarity in page-based multimodal text. *JASFL Occasional Papers, 1*(1), 25–49.

Teedaaksornsakul, M. (2024). Design and development of intersemiotically cohesive teaching materials for English for agriculture [Doctoral dissertation, Thammasat University]. TU Digital Collections. https://ethesisarchive.library.tu.ac.th/thesis/2023/TU_2023_6206320027_18820_28073.pdf

Teedaaksornsakul, M., & E. J. A. Bowen, N. (2024). Students' perceptions of intersemiotically cohesive teaching materials designed for English for agriculture. *rEFLections, 31*(2), 568–591. https://doi.org/10.61508/refl.v31i2.274907

Van Leeuwen, T. (1991). Conjunctive structure in documentary film and television. *Continuum: Journal of Media & Cultural Studies, 5*(1), 76–114. https://doi.org/10.1080/10304319109388216

Yanuar, R., Tandiana, S.T., Hidayati, A.N., Andriani, A., & Saputra, Y. (2021). Representing Intersemiosis in an English language teaching textbook: A systemic functional multimodal discourse approach. *Koli Journal 2*(1), 18–34. https://doi.org.10.30598/koli.v2i1.5120