The Journal of Language and Literature Insight
Volume 2, Issue 3, P. 17 – 25
e-ISSN: 3048 - 3441

# ANALYZING MULTIMODALITY IN ELT VIDEOS

**Novin Berliana** ✉
English Education Department
UIN Raden Intan Lampung
Indonesia

**Mohammad Muhassin**
English Education Department
UIN Raden Intan Lampung
Indonesia

**Satria Adi Pradana**
English Education Department
UIN Raden Intan Lampung
Indonesia

**Dewi Ayu Hidayati**
Sociology Department
Universitas Lampung
Indonesia

**Article Information**

Received: September 30, 2025
Revised: November 18, 2025
Accepted: December 9, 2025

**Abstract**

This study aims to investigate how the visual and verbal signs work together in constructing meaning in instructional videos. This is a descriptive qualitative research study. The data were derived from ELT videos in YouTube titled "Teaching 21st Century Skills: Oxford Discover Sample Lesson Level 3" (parts 1-4). The analysis follows the framework of Multimodality, Visual Grammar, and Intersemiotic Complementarity. The findings cover several modes used in the videos: linguistic, visual, gestural, and spatial. These modes are used in a combination of two or more modes in one video frame entitled "Teaching 21st Century Skills: Oxford Discover Sample Lesson Level 3". In addition, this study identified narrative processes, including action and reaction, eye gaze, distance, camera angles, information value, salience, and framing. These elements are used in a frame, combined with mode or multimodality, and are connected to create meaning. In particular, the verbal and visual modes support each other to form cohesion through the relation of repetition, synonyms, homonyms, and meronyms. The combination of these modes conveys a message to the audience, making it easier to understand the video's content as a whole. These findings imply that instructional video designers and EFL educators should strategically integrate visual and verbal modes to enhance clarity and learner engagement.

**Keywords:** Multimodality; Instructional Videos; Intersemiotic

Corresponding author: berliananovin@gmail.com

**Introduction**

Language is a human communication tool in conveying ideas, feelings, and statements. As stated by Hybel & Weaver (2004), who define "communication as a process of interaction between one person and another to share information and beliefs, exchange thoughts and feelings, make plans and solve problems". In other words, humans cannot be separated from language because, in every activity, especially in communicating with each other, they need it as an essential part of their daily lives.

Nowadays, multimodal communication has surpassed monomodal communication as the preferred mode of human communication. Most magazines and newspapers employ visual imagery to assist readers in understanding the topics and contents of articles. Photographs, caricatures, and cartoons are examples of visual imagery used in newspaper and magazine articles to convey information. Multimodality is a characteristic in these discourses. As stated by Lim (2002) “We live in a multimodal society, which produces meaning through the co-employment of semiotic resources”.

In a linguistics-based perspective, verbal and visual elements are essential for the multimodal approach. Multimodal analysis, as Kress and Van Leeuwen (1996) argued, has developed into new research areas that can commonly be applied in various fields of study, from literature, art, classroom teaching, and other fields of study. The term multimodality refers to the use of multiple modes of communication and representation to encourage the audience to pay more attention to aspects beyond the text when sharing their thoughts, such as visual modes like images, color, typography, and composition (Hidayat, 2015). One example of multimodality is YouTube videos. They are an example of multimodality, which consists of verbal and visual elements. Multimodality refers to a combination of writing, speaking, visualization, sound, music, and more. Literacy usually refers to the combination of letters, words, and pictures that leads to a conclusion about a message (Kress & van Leeuwen, 2010).

Usually, multimodal texts aim to perform communicative functions through verbal and semiotic modes. Multimodality deals with the communication and representation of thoughts in other modes (Kress & van Leeuwen, 2001). It needs to be able to convey the idea or thoughts to the audience without a thorough explanation. Kress and Leeuwen (1996) investigate the details of the concept of the multimodal perspective to uncover its underlying meaning. It deals with movement, image, sound, speech, and music. These are all the bases of interpersonal meaning, which are the concepts substituted and delivered by the meaning resources. The most investigated aspects are visual and verbal texts (Kress & van Leeuwen, 2001). Multimodality involves dealing with ideational functions that express the speaker's thoughts and with interpersonal functions that use not only language but also visuals. Both have a role relationship in the unitary text (Dornyei, Z, 2007). Based on the theory, multimodality can be explained through its various objects and their relations, which together represent the intended meaning as an intact unit.

Previous researchers have also studied multimodality with a different focus. First, the study conducted by Muhassin in 2022, titled "A Multimodal Analysis of Umrah Pilgrimage Advertisement". This research aims to explore the visual and verbal modes employed in Umrah pilgrimage advertisements, the meanings conveyed by these modes, and the meaning relations they build to strengthen the advertisement message. Utilizing the generic structure framework of advertising, visual data were analyzed by Visual Grammar Kress & van Leeuwen, and verbal data were scrutinized by Systemic Functional Grammar by Halliday & Matthiessen. Meanwhile, the intermodal meaning relations were viewed through Royce's Intersemiotic Complementarity. The research found that the advertisement had all the generic structures, namely lead, display, visual emblem, announcements, enhancer, verbal emblem, tag, and call-and-visit information. In verbal modes, the nominal groups represent the ideational meaning, whereas the imperative clause serves as ideational, interpersonal, dan textual meanings (Kaźmierczak, 2018). In addition, visual modes include representational, interactive, and compositional meanings. Verbal and visual modes support each other to form cohesion through the relations of repetition, synonymy, homonymy, meronymy, and collocation. With cohesion, the advertisement's message becomes

more communicative and persuasive, thereby arousing readers' interest in the company's product services.

The second research by Kassandra, Sinar, and Zein in 2018 applied the theory of Multimodal Interactive Meaning in their research to find out the meaning of interaction contained in a video game in the form of visual images and also the message contained in the video game that is found in the form of representation or adaptation. The game, Defense of the Ancients 2, or more widely known as DOTA 2, is one of the most popular games in society, especially amongst gamers. Many characters in this video game are inspired by real-world myths and legends in the real world, which is also the main reason the writer chose the game as an object. In this thesis, Kassandra applied the method Interactive Model from Miles, Huberman and Saldana (2014) and Jewitt's multimodal theory to analyze video games through images displayed on the screen on the meaning of interaction. The thesis examined the visual components found in the loading screens for characters in the video game and the similarities between the adaptation and myths.

The result of the research showed findings about interactive meanings that are applied on images in the video game and the similarities and differences between the visual images of the characters in the game with the characters or creatures in the real myths and legends in terms of adaptation. The characters' skills were also related in the analysis of their adaptation.

The last research by Nadiyah in 2019 was titled "Multimodal Analysis of English Textbook Tenth Graders Senior High School (A Case of Buku Bahasa Inggris Kelas X by Indonesian Ministry of Education and Culture)". This study aimed to analyze how multimodality in an English textbook for Tenth Graders of Senior High School may scaffold learning through visual texts. It concerns on multimodal analysis of images, integrated with verbal texts and proposed language activities to explain how the visual meanings may enhance students' understanding of language and content. The findings showed the visual images and verbal text combine with each other to expose visual meanings. In addition, the images contribute to scaffolded learning in being part of the overall meaning (Dusi, 2015). Therefore, the researcher concluded that the images support students' understanding of the activities in the textbook.

The gap between related studies and this study is that, although the theories are grounded in multimodal visuality by Kress & van Leeuwen's and Halliday's SFG, this study analyzed English Language Teaching in YouTube Videos as the data. Not only is it not analyzed much, but it is also interesting because YouTube is one of the most used social media platforms with billions of users. The importance of analyzing this study is that YouTube has had such an impact on our generation. Everybody watches YouTube, and through YouTube, we can all also learn new things. YouTube has become an essential aspect in life in terms of entertaining and/or conveying thoughts to the audience. This is the reason why they created the video.

## Methods

This study employed a qualitative approach in the framework of multimodal analysis of the English Language Teaching videos. The data in this study are in the "Teaching 21st Century Skills: Oxford Discover Sample Lesson Level 3" (part 1-4) videos. The data were taken from the English Language Teaching videos downloaded from YouTube. Data analysis was carried out in an eclectic method by applying the theories of Systemic Functional Grammar (Halliday, 2004; Halliday & Matthiessen, 2014), Visual Grammar (Kress & van Leeuwen, 2006), and Intersemiotic Complementarity (Royce, 2007).

## Findings and Discussion

The analysis focuses on describing modes of multimodality from the video. The second describes how those modes present the message on the "Teaching 21st Century Skills: Oxford Discover Sample Lesson Level 3" videos through verbal mode; including transitivity systems, followed by visual analysis, and intersemiosis between visual and verbal modes.

Figure 1. Video of English Language Teaching Part 1

This video is one of those from Oxford University Press, a channel of the international English teacher community that releases videos on teacher trainers and offers advice and tips to improve English teaching. The video chosen in this study is an English teaching video entitled "Teaching 21st Century Skills: Oxford Discover Sample Lesson Level 3". The Video was taught by a teacher named Charles Vilina. In this video, the course develops students' 21st-century skills and their English in they deliver lessons to enhance Communication, Collaboration, Creativity, and Critical Thinking through a variety of structured class activities from Oxford Discover.

***Multimodality used in English Language Teaching Videos***

***1. Linguistic Analysis***

Linguistics in multimodality covered oral and written language, including the text's generic structure, vocabulary, and grammar. Based on the findings, the ELT videos entitled “Teaching 21st Century Skills: Oxford Discover Sample Lesson Level 3” (part 1 - part 4) contain linguistic modes such as written text and oral text. The written text includes the learning steps. While for oral text, include the teacher's English language expression when having interaction with the students.

***2. Visual Analysis***

According to Anstey and Bull, the visual focus on how several modes in the visual aspect will contribute to creating meaning. In his findings, the visual mode in the ELT videos entitled “Teaching 21st Century Skills: Oxford Discover Sample Lesson Level 3” (part 1 - part 4) shows all the information and activities of the teaching and learning process, starting from the description of topics, classes, and teaching methods. Participants in the video consist of teachers and students who interact with each other. Other visual modes in the video include poster visuals, worksheet visuals, and plant video visuals.

***3. Gesture Analysis***

Gesture mode describes each participant's movements, body language, and facial expressions. Based on the findings of this study, in the ELT video series “Teaching 21st Century Skills: Oxford Discover Sample Lesson Level 3” (part 1-4), both teachers and students are involved in all aspects of gesture mode. For movement, the teacher sometimes goes around the class to check student work and also answers or explains more if students don't really understand what they have to do.

Next is body language. When the teacher explains the material, they often use body language, such as hand movements and eye contact with students. For facial expressions, the teacher looks very enthusiastic when delivering material and explaining what students are doing about their individual or group work, while students are very enthusiastic about listening to the teacher's explanation and working on worksheets and group work.

***4. Spatial Analysis***

Based on research findings on spatial analysis in ELT videos entitled "Teaching 21st Century Skills: Oxford Discover Sample Lesson Level 3" (part 1-4), the distance between the teacher and students indicates a multimodal spatial mode. According to Anstey and Bull, spatial mode also uses participants to analyze it. The teacher's position is in front, while the students are behind and pass each other. The teacher, as the center of attention of students, looks very enthusiastic in explaining the material. A study by Suprakisno found that spatial analysis of Indomie advertisements involves measuring the distance between images, where each image has its own meaning. However, the pictures support each other's meaning. In contrast to this study, the spatial analysis includes the distance between the teacher and students.

***Verbal Mode***

This study examined the verbal mode of the ELT video series "Teaching 21st Century Skills: Oxford Discover Sample Lesson Level 3" (Parts 1–4) through the lens of the transitivity system in Systemic Functional Linguistics (SFL), focusing on Process Types, Participants, and Circumstantial Elements. The analysis revealed 281 clauses, 491 participants, and 106 circumstantial elements, indicating a rich use of linguistic resources to construct meaning and facilitate interaction in the instructional context.

The findings demonstrate that all six process types proposed by Halliday are present in the videos: material, mental, relational, verbal, behavioral, and existential. Material processes appear to be the most prominent, reflecting the instructional nature of the videos, which frequently emphasize actions and activities performed by students, such as eating, working, moving, and sharing. This dominance suggests that the videos prioritize experiential learning and physical engagement, aligning with communicative and task-based approaches commonly applied in 21st-century skill instruction.

Mental processes also occur frequently, indicating the teacher's efforts to stimulate learners' cognitive engagement by encouraging them to remember, see, know, and think. Such processes position students as active thinkers and reflect an emphasis on internal cognitive processes, which is essential in promoting higher-order thinking skills. Relational processes are used to describe and evaluate concepts, as seen in expressions that assign attributes or identities, thereby supporting explanation and clarification in teaching.

Verbal processes further highlight the interactive nature of the classroom discourse, as the teacher frequently asks questions, gives instructions, and elicits responses. This pattern illustrates how spoken interaction is utilized to maintain engagement and guide learners through the lesson. Behavioral and existential processes, although less frequent, contribute to portraying physical reactions and the existence of objects or concepts within the learning environment, reinforcing the realism and contextual grounding of the instructional content.

The high number of participants reflects the complexity of clause construction, particularly in material and mental processes where actors, goals, sensers, and phenomena are clearly defined. This indicates that the videos consistently identify who is involved in each action or mental activity, thereby clarifying meaning and reducing ambiguity for learners. The prominent use of participants in material and verbal processes also highlights the focus on interaction between the teacher and students.

Circumstantial elements further enrich the meaning by providing additional information related to location, manner, extent, cause, accompaniment, matter, and rolem (Dweik & Suleiman, 2013). The frequent use of manner and location circumstances suggests that the videos aim to specify how and where actions take place, enhancing clarity and contextual understanding. Meanwhile, the cause-and-purpose circumstances demonstrate that explanations are often connected to intended outcomes, supporting logical reasoning and comprehension.

Overall, the transitivity patterns reveal that the ELT videos employ a balanced combination of action-oriented, cognitive, and interactive language. This strategic use of verbal processes supports effective knowledge transmission and fosters learner engagement by connecting physical activity with cognitive processing, ultimately contributing to a more dynamic and meaningful learning experience.

***Visual Mode***

This learning activity on this figure is contained in the video, which runs from 0.58 – 11.20 minutes. This figure was an example of Transactional Action in Ideational Metafunction. In this figure, the teacher was explaining about a plant. The teacher showed a sunflower and asked the students to look at it and pay close attention. Then the teacher tried to make the students aware of what topics they would learn. After that, the teacher asked some questions to the students to help them identify what they would like to know about the subject, like "is it a plant or an animal, what kind of plant is it, and so on. Then the student said it was a plant and that it was a sunflower. In this scene, the actor was the teacher, and the goal is the students. The students need to be goal-oriented when listening to the teacher's explanation. The goal was for the participant who receives the action (Kress & van Leeuwen, 2006). While the teacher is the actor because she becomes more prominent than the students. It was examined from the camera shots to the teachers' whole faces.


Figure 2. The teacher Explained about the Plant

The visual imagery of the elements of the Interpersonal Metafunction/Interactive meaning included in this learning activity includes gaze, distance, and camera angle. The first element is gaze. The gaze of this figure is to offer gaze. Offer gaze refers to participants' gaze on the video, which indirectly targets viewers (Kress & van Leeuwen, 2006). The following element is social distance. The social distance in this video is evident in the frame size. The distance is indicated to the distance between the participants on the video with the viewers. According to Kress and Van Leeuwen's theory (2006), there are three kinds of distance: long shot, medium shot, and close-up. In this learning activity, all activities use the same frame size: medium shot. The medium shot cuts roughly at the subject's waist or knee. It showed that the participant in that figure was only visible from the knee down, not the whole body. Moreover, the camera angle is an eye-level shot. The shot is taken in which indicates that the distance between the subject and the viewer is equal. There is no difference in strength between the participants in the video with the viewer.

Compositions, that is, the representational and interactive meanings are combined into a unity of significance. The compositional/Textual metafunction is achieved through three interrelated systems: information value, salience, and framing (Kress van Leuween, 2006). In this study, we will take one of the scenes that expound the learning activity in ELT videos, teaching preparation as an example to analyze the compositional meaning in the video.

Firstly, for information value, each object or participant is put in a different placement in each scene of the video. Thus, placement affects the participants' importance. There are several compositional placements based on how it contains an information value to the audience. First, the left-to-right composition gives a value of ‘given’ to ‘new’ information. Second, the top-to-bottom composition provides a value of ‘ideal’ or ‘unreal’ to ‘real’ information. Third, the

composition of centre and margin gives an information about the importance of the object that is placed in the central part of the image.

In Silence, salience is the elements that attract the audience's attention. it shows that the salience goes to the teacher. The teacher is the most salient object or participant in the video. It can be seen from how the camera shot is mainly focused on the teacher's movement. Another supporting element that shows the teacher as the most salient object is in the camera shot. The teacher was primarily shot in medium-close and long shots to show the main focus during the scenes. The teacher was also placed at the very front, while the other participant was usually placed in behind her or beside him.

The last was Framing. Framing can be found when there are dividing lines or actual screen lines to move to another scene. The transition usually separates the latest scene from the newest one. If each element is connected to the others, it will convey a strong message or information to the audience.

***Intersemiosis between the Visual and Verbal Modes***

This part discusses the meaning relations between the visual and verbal modes of the ELT videos. In the video, several attempts were made to harmonize the meanings of the two modes in the ELT videos. The combination of visual and verbal modes in multimodal texts will contribute to the achievement of intersemiotic complementarity. Based on the analysis, this study found the use of intersemiotic complementarity, including intersemiotic repetition, synonymy, hyponymy, and meronymy.

***Intersemiotic Repetition***


Figure 3. Teacher Holding a Sunflower

Figure 4. Teacher pointing to a picture in a book

Intersemiotic repetition occurs when the meaning of two semiotic modes is the same (royce 2007). The intersemiotic relation is seen by relating the visual and verbal modes. From figures 7 and 8, it is clear that the Sunflower and the picture in the book are portrayed in both multimodal modes, thereby repeating intersemiotic relations. The verbal and visual elements can be captured in the intersemiotic relation. In this finding, the repetition in a visual mode with the same meaning makes it easier for viewers to recognize the lexical item 'it is sunflower' in figure 7, and the lexical 'guys let's look at the picture'.

***Hyponym***

One instance of intersemiotic hyponymy was observed in the English Language Teaching video, part 1. A visual showing a teacher holding a sunflower, accompanied by a verbal statement referring to the lexical item “is there anything the same between plants and animals?” So, in that instance, the verbal mode indicated the general class and the visual mode showed the sub-class. The visual presentation conveys the general meaning that a sunflower is one of the various types of plants. In the context of hyponymy relations, visuals containing images of sunflowers are classified as a subclass of lexical items of plants. This finding supports Royce's argument that

intersemiotic hyponymy can build cohesiveness of meaning in multimodal texts by linking general class meanings to their subclasses.


Figure 5. The Teacher Tells about the plant

***Synonym***

Furthermore, the case of intersemiotic synonyms appears in the visual, which contains a teacher pointing to the diagram in the whiteboard. The visual presentation, supported by the lexical item "two ways that plants and animals are the same and two ways that plants and animals are different," showed two things with the same meaning. Both the verbal and visual modes complement each other to help learners understand the similarities and differences between plants and animals through diagrams.

***Meronym***

Meronymy: intersemiotic findings in the ELT videos. When in visual mode, the teacher shows a picture in the book of children collecting various kinds of plants. It implies an lexical item "these are children gathering vegetables" as part of a visual "children collecting kinds of plants" referring to the whole. This shows that there is an overall relationship between the visual and verbal modes. that meronymy is helpful because it facilitates the learning of part-whole relationships in multimodal texts. In this context, the visual mode, which depicts children collecting various kinds of plants, conveys the overall meaning, and this meaning is reinforced by its members, namely the lexical children who are collecting vegetables.

## Conclusion

The study has found the use of several verbal and visual modes in ELT videos, including linguistic, visual, gesture, and Spatial modes. The verbal elements are represented through clauses that consist of experiential functions, namely Material, Mental, Behavioural, Existent, Verbal, and Relational Processes. The participants are Actor, Goal, Scope, Receiver, Initiator, Senser, Phenomenon, Carrier, Attribute, Identified, Identifier, Sayer, Target, Receiver, Verbiage, Existent, Behaver, and Behavior, completed by the Circumstances. Meanwhile, the visual modes representational, interactive, and compositional meaning.

This study has also identified the use of intersemiotic complementarity, including intersemiotic repetition, synonymy, hyponymy, and meronymy. By combining various modes of communication, the ELT video appeared more interesting, lively, and easy to understand for the viewers. Further research is recommended to broaden the topic of ELT videos with the analysis of semiotics and pragmatics to complete the limitations of this study.

## References

Anstey, M. & Bull, G. (2010). Helping teachers to explore multimodal texts. *Curriculum and*

*Leadership Journal. 8(16).*

Dusi, N. (2015). *Intersemiotic translation: Theories, problems, analysis*. *2015*(206), 181–205. https://doi.org/10.1515/sem-2015-0018

Dweik, B. S., & Suleiman, M. Y. I. H. (2013). Problems Encountered in Translating Cultural Expressions From Arabic Into English. *International Journal of English Linguistics*, *3*(5). https://doi.org/10.5539/ijel.v3n5p47

Halliday, M. A. K. & Matthiessen, C. M. I. M. (2014). *Halliday's introduction to functional grammar* (4th ed), New York: Routledge.

Hidayat, A. (2015). Content Analysis of the Lexical Density of the English for Islamic Studies Textbook of Iain Raden Intan Lampung. *English Education: Jurnal Tadris Bahasa Inggris IAIN Raden Intan*, *8*(1), 119–138. https://doi.org/10.24042/ee-jtbi.v8i1.513

Kaźmierczak, M. (2018). *From Intersemiotic Translation to Intersemiotic Aspects of Translation*. *Numery anglojęzyczne*(Special Issue 2018 – Word and Image in Translation), 7–35. https://doi.org/10.4467/16891864ePC.18.009.9831

Kress, G. R. & van Leeuwen. (2006). *Reading Images: The Grammar of Visual Design.* 2. ed., reprinted. London: Routledge.

Muhassin, M. (2022). A Multimodal Analysis of Umrah Pilgrimage Advertisement. *Jurnal Education and Development, 10*(1), 460-469.

Muhassin, M. (2023). Transitivity and Modality Analysis of Tedros Adhanom Ghebreyesus's Speeches on Handling COVID-19. *Theory and Practice in Language Studies, 13*(6), 1581-1590.

Nadiyah. (2019). Multimodal Analysis of English Textbook Tenth Graders Senior High School (A Case of Buku Bahasa Inggris Kelas X by Indonesian Ministry of Education and Culture).

Royce, T. (2002). Multimodality in the TESOL Classroom: Exploring Visual-Verbal Synergy. *TESOL QUARTERLY, 36* (2).

Royce, T. (2007). Intersemiotic complementarity: A framework for multimodal. In T. D. Royce & W. L. Bowcher (Eds.), *New directions in the analysis of multimodal discourse* (pp. 63-109). London: Lawrence Erlbaum Associates