KAMPUS AKADEMIK PUBLISING
Jurnal Multidisiplin Ilmu Akademik (JMIA)
Vol.3, No.3 Juni 2026
e-ISSN: 3032-7377, p-ISSN : 3032-7385, Hal 235-244
DOI : https://doi.org/10.61722/jmia.v3i3.10048

CONSTRUCTING MEANING THROUGH MULTIMODAL SEMIOTICS: AN ANALYSIS OF MS. RACHEL'S VIDEO *BABY LEARNING WITH MS. RACHEL - FIRST WORDS, SONGS AND NURSERY RHYMES FOR BABIES – TODDLER VIDEOS*

Fauzya Saputra[1]*, Sony Junaedi[2]
[1]*Faculty of Languages and Culture, Universitas 17 Agustus 1945 Semarang, Indonesia*
[2]*Faculty of Languages and Culture, Universitas 17 Agustus 1945 Semarang, Indonesia*
*Penulis Korespondensi: fauzyasaputra@gmail.com , sony-junaedi@untagsmg.ac.id

Abstract. *This qualitative descriptive study explores how meaning is constructed through multimodal semiotics in Ms. Rachel's YouTube video 'Baby Learning – First Words, Songs and Nursery Rhymes for Babies – Toddler Videos.' Grounded in Kress and van Leeuwen's (2001, 2006) multimodal semiotics framework and supported by Norris's (2019) multimodal interaction analysis, the study analyzes three selected video segments to examine how visual, verbal, gestural, and audio-musical modes interact to support early language acquisition. Analysis of key pedagogical moments reveals five dominant multimodal strategies: multimodal redundancy, temporal synchronization, functional complementarity, progressive modal complexity, and compensatory modal distribution. The findings demonstrate that meaning is constructed not merely through the presence of multiple modes, but through their strategic temporal coordination and functional specialization. Visual elements provide concrete representations, verbal elements facilitate phonological development, gestural elements reinforce meaning through iconic actions, and audio-musical elements structure the learning flow and emotional engagement. The study contributes theoretically by highlighting the importance of temporal synchronization in multimodal meaning-making and practically by providing evidence-based guidance for educators and content developers designing early childhood learning media. Results indicate that effective multimodal learning requires careful orchestration of semiotic resources across time, not merely their simultaneous inclusion..*

Keywords: *multimodal semiotics, multimodal interaction analysis, early childhood language learning, digital learning media, semiotic affordance*

1. INTRODUCTION

In the digital age, educational video content has become an integral component of early childhood language learning worldwide. Platforms such as YouTube have transformed how young children access learning resources, providing dynamic combinations of visual, auditory, and interactive stimuli that support cognitive and linguistic development (Rohmah & Aziz, 2024). Among the most influential creators in this space, Ms. Rachel's 'Baby Learning' series has garnered millions of subscribers and billions of views globally, becoming a prominent model for research into how multimodal resources can effectively facilitate early language acquisition. From a theoretical perspective, multimodal social semiotics provides a robust framework for understanding

Naskah Masuk: 12 Mei 2026; Revisi: 13 Mei 2026; Diterima: 12 Mei 2026; Terbit: 14 Mei 2026

how meaning emerges not through language alone, but through the coordinated interaction of multiple semiotic modes working together (Kress & van Leeuwen, 2006).

Early childhood learning relies fundamentally on the integration of visual, auditory, and motor inputs to develop understanding and communicative competence (Heinrich et al., 2020). While previous research has documented the effectiveness of multimodal approaches in classroom settings, fewer studies have examined how meaning is constructed through multimodal semiotics in popular digital learning media accessed by millions of families globally. Existing studies have either emphasized quantitative impacts of media consumption or conducted structural analysis without accounting for temporal synchronization of modes, both of which remain underexplored in the context of popular children's educational videos. This gap in literature motivated the present study: to systematically analyze how multiple semiotic modes interact temporally and functionally in a globally influential educational video to construct meaning that supports early language acquisition.

The present study addresses two primary research questions: (1) What are the multimodal semiotic resources employed in Ms. Rachel's video, and how does each mode contribute to meaning construction? (2) How do these modes interact temporally and functionally to support early childhood language acquisition? To answer these questions, the study applies an integrated analytical framework combining Multimodal Semiotics (Kress & van Leeuwen, 2001, 2006) as the primary theoretical lens with Multimodal Interaction Analysis (Norris, 2019) as a supporting approach. This combination enables analysis at both macro and micro levels, capturing both the structural roles of different modes and their real-time coordination. The novelty of this research lies in its examination of temporal synchronization—an often-overlooked dimension in multimodal analysis—and its application to popular children's digital learning content with significant real-world educational impact.

2. LITERATURE REVIEW

Multimodal semiotics, as developed by Kress and van Leeuwen (2006), posits that communication involves multiple semiotic resources—visual, verbal, gestural, and audio-musical—each operating through three metafunctions: ideational (representing objects and concepts), interpersonal (establishing relationships with audiences), and textual

(organizing information sequentially). The concept of semiotic affordance is central to this framework, stipulating that each mode has distinct meaning-making capacities and limitations (Kress & van Leeuwen, 2001). For instance, the visual mode effectively represents spatial relationships and concrete objects, while the verbal mode excels at providing linguistic labels and abstract concepts. The gestural mode communicates through movement and action, and the audio-musical mode structures temporal flow and emotional experience.

In the context of early childhood education, multimodal approaches have demonstrated strong pedagogical effectiveness. Heinrich et al. (2020) demonstrated that crossmodal integration—the coordinated processing of visual, auditory, and sensorimotor information—is fundamental to language learning. Deklerk (2020) found that multimodal interactions encompassing gestures, movement, and music support meaning-making in young children more effectively than single-mode instruction. Furthermore, Samuelsson (2023) showed that exposure to diverse multimodal environments enhances children's literacy development and cognitive flexibility. These findings collectively suggest that children benefit from learning environments where information is presented through multiple coordinated modes rather than linguistic input alone.

YouTube has emerged as a significant platform for early childhood learning, with both pedagogical value and potential risks. Radesky et al. (2022) noted that video-sharing platforms support children's language and cognitive development through engaging multimodal exposure, though they emphasize the necessity of parental guidance to ensure content quality. Content creators like Ms. Rachel have gained prominence precisely because their videos integrate research-based pedagogical elements such as infant-directed speech (parentese), repetition, gestural scaffolding, and direct address—all well-established features supporting early language acquisition (Friska, 2025).

However, most existing research addresses either the quantitative impacts of media consumption or provides descriptive analyses of multimodality without examining the temporal dimension—how modes interact in real-time sequences. Norris (2019) introduced Multimodal Interaction Analysis to address this gap, emphasizing concepts of modal intensity (how concentrated multiple modes are in brief moments) and temporal coordination (how modes align or diverge across time). This temporal perspective

remains underutilized in educational media analysis but is crucial for understanding how meaning emerges dynamically in video-based learning. By integrating Kress and van Leeuwen's framework with Norris's temporal emphasis, the present study fills this theoretical gap while providing insights applicable to millions of families using YouTube for early childhood education

3. RESEARCH METHOD

This qualitative descriptive study employs multimodal semiotics analysis to examine Ms. Rachel's video 'Baby Learning with Ms. Rachel – First Words, Songs and Nursery Rhymes for Babies – Toddler Videos' (uploaded March 2, 2022; duration 01:00:21). The video was selected because it represents a comprehensive pedagogical framework combining multiple teaching modalities (songs, gestures, visual aids, direct address) specifically designed for early language acquisition. Data were collected through systematic observation and multimodal transcription, with analysis conducted using an integrated framework combining Multimodal Semiotics (Kress & van Leeuwen, 2006) and Multimodal Interaction Analysis (Norris, 2019).

Three representative video segments were selected for detailed analysis based on three criteria: diversity of vocabulary types (concrete vs. abstract), variety of teaching approaches (structured instruction, song-based, embodied movement), and progression in complexity (simple single words to complex emotional vocabulary). Segment 1 (00:00–03:48) introduces 'mama' and 'dada' through a five-stage sequence (say, sing, clap, sign, praise). Segment 5 (13:27–19:04) teaches animal vocabulary through 'Old MacDonald Had a Farm' using props and onomatopoeia. Segment 10 (39:33–43:24) addresses emotional vocabulary ('happy,' 'mad,' 'scared,' 'silly') through embodied movement and facial expression. These three segments exemplify the full range of multimodal strategies employed throughout the entire video.

Analysis proceeded through four sequential stages: (1) detailed multimodal transcription documenting verbal utterances, visual elements (camera framing, color, on-screen objects), gestures (pointing, hand movements, facial expressions), and audio-musical features (intonation, rhythm, melody); (2) identification of multimodal resources within four primary modes (visual, verbal, gestural, audio-musical); (3) classification of child-directed speech features (repetition, exaggerated prosody, simplified vocabulary,

parentese); and (4) interpretation of how identified resources interact temporally and functionally to construct meaning. Temporal analysis examined simultaneous and sequential mode interaction, while functional analysis identified the specific role of each mode in constructing ideational, interpersonal, and textual meaning. To ensure validity, interpretations were cross-checked with theoretical frameworks and previous research. Trustworthiness was established using the four criteria of Lincoln and Guba (1985): credibility, dependability, confirmability, and transferability.

4. RESULT AND DISCUSSION

4.1 Multimodal Resources and Their Functional Roles

Analysis reveals that Ms. Rachel's video systematically employs four distinct semiotic modes, each performing specialized functions in meaning construction. The visual mode encompasses gaze direction, color choices, camera framing, and on-screen objects. In vocabulary instruction, the visual mode functions primarily through the ideational metafunction by providing concrete, directly observable representations (e.g., a photograph of a mother accompanying the word 'mama'). The choice of close-up camera framing combined with warm colors creates an interpersonal connection simulating direct interaction between the speaker and child audience. This design reflects an understanding of how young children's attention and engagement are shaped by visual salience and social proximity cues.

The verbal mode encompasses spoken language features including word choice, repetition patterns, intonation contours, pace, and paralinguistic features. Ms. Rachel's verbal delivery is distinctly characterized by infant-directed speech (parentese), featuring exaggerated intonation, slower tempo with clear articulation, and melodic quality ('Can you say ma-ma? Maaa... maaa... good job!'). This verbal scaffolding serves both ideational functions (providing linguistic labels) and interpersonal functions (establishing emotional warmth and encouraging participation). The repetition of key vocabulary supports phonological awareness and word recognition through multiple exposures with prosodic variation, preventing monotony while reinforcing key information.

The gestural mode includes pointing gestures, hand movements, facial expressions, and body positioning. In the analysis, gestures function through two primary mechanisms:

iconicity and convention. Iconic gestures directly resemble their referents (e.g., making a snout shape with hands to represent a pig). Conventional gestures follow established sign language or cultural norms (e.g., American Sign Language signs for body parts). These gestural elements serve ideational functions by representing object characteristics and interpersonal functions by inviting children's motor participation through imitation. The coordination of gestures with speech—such as making animal sounds while performing the corresponding gestural representation—enables cross-modal meaning reinforcement.

The audio-musical mode comprises background music, melodic patterns, rhythm, and sound effects. Rather than serving purely decorative functions, music in Ms. Rachel's videos performs critical pedagogical roles. Repetitive musical patterns organize information into predictable chunks, reducing cognitive load (a principle well-established in learning science). Rhythm aligns with syllable structure (e.g., hand clapping synchronized with 'ma-ma'), facilitating phonological segmentation. Melodic variation creates emotional engagement and enhances memory retention through strong auditory-affective pathways. This orchestration of musical elements reflects understanding of how rhythm and melody support temporal organization of information and emotional investment in learning.

4.2 Patterns of Multimodal Integration

Analysis identifies five dominant patterns in how modes interact to construct meaning: (1) Multimodal redundancy: The same concept is represented through multiple modes, but not identically. For example, 'mama' is conveyed through a verbal label, a photograph, an ASL sign, and accompanying music. Each mode presents a different representational perspective, leveraging each mode's specific affordances to enrich understanding rather than creating wasteful duplication. (2) Temporal synchronization: At pedagogically significant moments, multiple modes concentrate simultaneously for brief periods. When Ms. Rachel introduces the word 'mama,' visual (photograph), verbal ('mama'), gestural (ASL sign), and musical elements align within approximately 20 seconds. This temporal coordination creates high modal intensity that functions to emphasize information and focus children's attention.

(3) Functional complementarity: Each mode contributes distinct information appropriate to its affordances. Concrete vocabulary learning relies more heavily on visual

representation, while emotional vocabulary requires greater emphasis on facial expression and prosodic variation. (4) Progressive modal complexity: The video demonstrates developmental progression from simple (Segment 1: single words with straightforward multimodal support) to more complex (Segment 10: abstract emotions requiring coordinated facial expression, prosody, gesture, and movement). This scaffolding respects children's developing cognitive and perceptual capacities. (5) Compensatory multimodality: When single modes prove inadequate—as with abstract concepts like emotions—multiple modes work together to represent what cannot be represented in any single mode. Children construct understanding of 'happy' by integrating facial expression, vocal enthusiasm, energetic movement, and fast-tempo music into a cohesive emotional concept.

4.3 Theoretical Contributions and Implications

These findings contribute to multimodal theory in several important ways. First, they demonstrate that temporal synchronization—the timing and coordination of modes across time—is as critical to meaning-making as the presence of multiple modes themselves. Kress and van Leeuwen's framework, developed primarily for analyzing static texts, does not explicitly account for temporal dynamics. By demonstrating that meaning-making depends on real-time coordination of modes, this study extends multimodal theory to more adequately describe video-based communication. The analysis shows that simply including multiple modes is insufficient; pedagogical effectiveness depends on deliberate orchestration of when modes appear and how they align.

Second, the findings expand understanding of multimodal relationships beyond pure complementarity to include compensatory patterns. When modes work together to represent information that no single mode can adequately convey—particularly in abstract domains—they function compensatorily rather than complementarily. This distinction is important for pedagogical design: instructors must choose different modal combinations depending on whether content is concrete or abstract. Third, the analysis confirms previous research emphasizing crossmodal integration (Heinrich et al., 2020) while demonstrating the specific mechanisms through which integration occurs in video-based media: temporal coordination at high-intensity moments, functional specialization according to modal affordances, and systematic progression in complexity.

4.4 Practical Implications for Educational Design

The findings carry significant practical implications for educators and educational media developers. For teachers and parents supporting early language development, the analysis suggests that multimodal communication—combining speech with gesture, visual aids, and musical or rhythmic elements—should be central rather than supplementary to instruction. Simple practices such as naming objects while pointing, using exaggerated facial expressions during instruction, and incorporating playful gestures or songs align with the evidence from Ms. Rachel's approach and require no special technology. The temporal synchronization patterns observed suggest that moments of intensive multimodal support are particularly valuable; teachers might strategically concentrate multiple modes when introducing key vocabulary rather than distributing them throughout instruction.

For digital media developers, the analysis provides evidence-based design principles: (1) Each mode should be used according to its functional capacity rather than for decorative purposes. (2) Critical information should be presented through temporal coordination of multiple modes rather than through any single mode. (3) Abstract concepts require more intensive multimodal integration than concrete vocabulary. (4) Cognitive load should be managed through predictable musical or rhythmic patterns, allowing children to focus on vocabulary acquisition. (5) Progression from simple to complex multimodal presentations should align with developmental expectations. These principles are neither expensive nor technologically demanding; they require primarily thoughtful pedagogical design. The global success of Ms. Rachel's channel suggests that audiences recognize and value this approach, making evidence-based multimodal design a competitive advantage for educational content creators.

5. CONCLUSION

This study examined how meaning is constructed through multimodal semiotic resources in Ms. Rachel's educational video using integrated frameworks of Multimodal Semiotics (Kress & van Leeuwen, 2001, 2006) and Multimodal Interaction Analysis (Norris, 2019). The analysis demonstrates that early childhood language learning in digital media depends on strategic coordination of visual, verbal, gestural, and audio-musical modes. Rather than functioning in isolation, these modes interact through

patterns of multimodal redundancy, temporal synchronization, functional complementarity, progressive complexity, and compensatory distribution to construct meaning that supports language acquisition.

A central insight emerging from this research is the critical importance of temporal synchronization—the deliberate alignment of multiple modes at pedagogically significant moments. This temporal dimension, often overlooked in multimodal analysis, proves essential to understanding how digital media facilitates learning. Children are most effectively engaged and information is most powerfully reinforced when multiple modes concentrate simultaneously, creating high modal intensity that focuses attention and supports cognitive integration.

The study contributes theoretically by extending multimodal frameworks to adequately describe video-based communication with explicit attention to temporal dynamics, and practically by providing evidence-based design principles for educators and media creators working with young children. As digital media continues expanding its role in early childhood education, understanding the specific mechanisms through which multimodal resources support language learning becomes increasingly important. This research provides both theoretical grounding and practical guidance for designing educational media that authentically supports children's development. Future research should extend this analysis across multiple videos, investigate children's actual responses to different multimodal designs, and examine whether these patterns transfer across cultural and linguistic contexts.

REFERENCES

Azhari, T. S., & Indah, R. N. (2025). Early childhood vocabulary acquisition through multimodal strategies: A semiotic study of Miss Rachel. Aulad: Journal on Early Childhood, 8(3), 1483–1492.

Baldry, A., & Thibault, P. J. (2006). Multimodal transcription and text analysis: A multimedia toolkit and coursebook. Equinox.

Deklerk, H. M. (2020). Multimodality and young children's meaning-making: Songs, movement and the development of language. Early Child Development and Care, 190(7), 1058–1070.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2018). The SAGE handbook of qualitative research (5th ed.). SAGE Publications.

Dressman, M. (2019). Multimodality and language learning. In The handbook of informal language learning (pp. 39–55).

Friska, Y. (2025). YouTube Kids on children's English communication skills: Parents' beliefs and attitudes. Journal of English Language Teaching and Applied Linguistics, 7(1), 45–56.

Halliday, M. A. K. (1978). Language as social semiotic: The social interpretation of language and meaning. Edward Arnold.

Heinrich, S., Bhatt, P., Bhatt, N., & Bhatt, A. (2020). Crossmodal language grounding in an embodied neurocognitive model. Frontiers in Neurorobotics, 14, 577755.

Isik, O. (2025). Qualitative research approaches and data collection methods: Understanding meaning and experience. Journal of Humanities and Education Development, 7(6), 19–32.

Jewitt, C. (Ed.). (2014). The Routledge handbook of multimodal analysis (2nd ed.). Routledge.

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

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

Lemke, J. L. (2021). Cognition, context, and learning: A social semiotic perspective. In Situated cognition (pp. 37–55). Routledge.

Li, D. (2018). Critical media literacy: A social semiotic analysis and multimodal discourse of corpocracy. International Journal of Education & the Arts, 19(16).

Lim, F. V., & Toh, W. (2020). Children's digital multimodal composing: Implications for learning and teaching. Learning, Media and Technology, 45(4), 422–432.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. SAGE Publications.

Miles, M. B., & Huberman, A. M. (1994). Qualitative data analysis: An expanded sourcebook (2nd ed.). SAGE Publications.

Norris, S. (2019). Systematically working with multimodal data: Research methods in multimodal discourse analysis. Wiley-Blackwell.

Radesky, J. S., Peacock-Chambers, E., Zuckerman, B., & Silverstein, M. (2022). Video-sharing platform viewing among preschool-aged children. JAMA Pediatrics, 176(4), 412–414.

Rohmah, B., & Aziz, T. (2024). Perkembangan bahasa anak usia dini di era digital: Dampak media YouTube, peran pengasuhan, dan perubahan sosial. Jurnal Warna: Pendidikan dan Pembelajaran Anak Usia Dini, 9(2), 213–229.

Samuelsson, R. (2023). Creating a web of multimodal resources: Examining early childhood literacy practices in a multilingual community. Journal of Early Childhood Literacy, 23(1), 3–28.

Sommer, V. (2021). Multimodal analysis in qualitative research: Extending grounded theory through the lens of social semiotics. Qualitative Inquiry, 27(8–9), 1102–1113.