ICRRD QUALITY INDEX RESEARCH JOURNAL
INTERNATIONAL CENTER FOR RESEARCH AND RESOURCES DEVELOPMENT article
ISSN Number: 2773-5958, https://doi.org/10.53272/icrrd, www.icrrd.com

From Screens to Semantics: A Qualitative Inquiry into Vocabulary Learning through Captioned YouTube Videos in Bangladeshi Higher Education

Sultana Jahan

Department of Language and Communication, Patuakhali Science and Technology University, Bangladesh

**Corresponding author; Email: sjkoli10@gmail.com*

Received: 15 March 2025 **Accepted:** 17 August, 2025
Revision: 29 July 2025 **Published:** 20 October 2025. **Vol**-6, **Issue**-3

Cite as: Jahan, S. (2025). From Screens to Semantics: A Qualitative Inquiry into Vocabulary Learning through Captioned YouTube Videos in Bangladeshi Higher Education. *ICRRD Journal, 6*(3), 321-338.

Abstract: This study explores how Bangladeshi university students develop English vocabulary through engagement with captioned YouTube videos, conceptualised at the intersection of incidental learning, multimodality, and connectivism. Grounded in the view that vocabulary acquisition emerges through meaning-focused, multimodal, and digitally networked encounters, the study examines how learners notice, process, and apply new lexical items in their informal viewing practices. Drawing on semi-structured interviews, reflective journals, and content-use logs from 15 students across two private and one public university, data were analysed through reflexive thematic analysis. Four interrelated themes captured learners' experiences: captions as adaptive scaffolds for noticing and recall; learning beyond classroom boundaries through curated digital routines; balancing entertainment and education to sustain motivation; and coping with infrastructural and cognitive constraints. Findings reveal that students transform captioned viewing into strategic, identity-driven vocabulary learning by managing caption modes, pacing, and genre selection. The study proposes integrating "caption literacy" and multimodal awareness into tertiary English curricula, recognising informal digital environments as legitimate learning spaces. By connecting cognitive, social, and technological dimensions of vocabulary acquisition, this research extends current understandings of informal digital learning and offers context-responsive insights for language pedagogy in the Global South.

Keywords: *Captioned YouTube videos; incidental vocabulary learning; multimodal pedagogy; informal digital learning; Bangladeshi higher education*

1. Introduction

Digital media and social platforms have fundamentally reshaped how learners encounter, process, and use language in the twenty-first century. Among these, YouTube stands out as one of the most influential informal learning spaces, offering authentic, multimodal exposure to English across diverse genres and registers. Bangladeshi university students, who often experience limited communicative

opportunities in traditional classrooms (Alam et el., 2025), captioned YouTube videos have become an accessible and engaging means of vocabulary learning. These videos integrate sound, text, and visual cues—allowing learners to map spoken forms onto written representations while simultaneously inferring meaning from context. Yet, despite their widespread use, the ways in which Bangladeshi students actually learn vocabulary through captioned YouTube content remain underexplored.

The Bangladeshi higher education system continues to grapple with persistent linguistic and pedagogical challenges (Milon, 2016; Al Nahar et al., 2024). English proficiency is widely viewed as essential for academic success and employability, but students' lexical competence often remains limited, particularly in productive use. Several studies note that many learners depend heavily on memorisation and grammar translation approaches, with minimal emphasis on communicative use or vocabulary recycling (Alam et al., 2018, 2021, 2025). This mismatch between curriculum design and real-world language use constrains learners' ability to internalise new lexical items effectively. While public universities face infrastructural barriers such as poor internet connectivity and large class sizes, private universities tend to have better access to digital resources but uneven pedagogical innovation (Hasan et al., 2019). Consequently, an increasing number of students turn to YouTube as a complementary space for exposure to authentic English, engaging with captioned videos to support listening, comprehension, and vocabulary growth. However, there is a lack of qualitative evidence on how learners in such contexts negotiate the opportunities and challenges of captioned video learning.

Globally, research has established that captioned videos enhance comprehension and vocabulary acquisition by synchronising auditory and textual input. Experimental studies have demonstrated that learners who view videos with same-language captions outperform those without captions in vocabulary recall and recognition tasks (Milon et al., 2018a, 2018b). Later studies have confirmed that captions promote noticing of new lexical forms and collocations, facilitating both incidental and intentional learning processes (Teng, 2019; Webb, 2020). Nevertheless, most of this evidence comes from well-resourced educational contexts, where learners enjoy stable technological access and teacher guidance. In contrast, less is known about how students in low-resource settings—such as those in South Asia—employ captions autonomously within their digital learning ecosystems. Understanding these situated practices requires moving beyond controlled experiments to explore learners' lived experiences, digital strategies, and adaptive behaviours in real-world environments.

To frame such experiences theoretically, this study draws on three complementary perspectives: incidental learning, multimodality, and connectivism. Incidental learning suggests that vocabulary can be acquired as a by-product of meaning-focused engagement, rather than through explicit memorisation (Nation, 2013; Milon & Ali, 2023). Multimodal learning theory argues that comprehension deepens when information is presented across visual and auditory channels, provided cognitive load is managed effectively (Mayer, 2020). In this light, captions serve as visual scaffolds that connect sound, orthography, and meaning. Connectivism (Siemens, 2005) extends this view into the digital age by positing that learning involves forming and sustaining connections across digital networks. Learners construct knowledge not only from individual texts or videos but also from links among playlists, comment sections, transcripts, and peer discussions. Integrating these frameworks allows us to conceptualise captioned video engagement as a networked multimodal process—one

where incidental noticing is activated through captions, reinforced by visual and auditory redundancy, and sustained through learners' active digital navigation and social participation.

Within this conceptual framework, the present study explores how Bangladeshi university students—across two private and one public institution—engage with captioned YouTube videos as tools for vocabulary development. It investigates what types of captions (English, bilingual, or autogenerated) and genres (academic lectures, interviews, explainers, entertainment) learners find most supportive, how they record and reuse new lexical items, and what contextual factors enable or hinder these practices. It also examines how vocabulary encountered in digital spaces is transferred to academic writing, presentations, and everyday communication.

To address these aims, a qualitative design grounded in reflexive thematic analysis (Braun & Clarke, 2019, 2021) was employed to uncover patterns of engagement, strategy, and meaning-making in participants' accounts. By privileging learner narratives and contextual detail, the study seeks to move beyond the question of whether captions "work" to explore how they facilitate vocabulary learning in digitally constrained settings. This interpretive focus highlights learners' agency, their negotiation of cognitive and infrastructural challenges, and the ways in which they transform informal media consumption into purposeful language learning.

The study makes three principal contributions. First, it integrates cognitive, sociocultural, and technological dimensions of vocabulary learning into a unified framework, addressing the fragmentation that characterises much of the existing literature. Second, it adds a Global South perspective to research on captioned videos and informal digital learning, documenting how learners in Bangladesh appropriate global technologies to meet local linguistic needs. Third, it offers concrete pedagogical insights by demonstrating how teachers and curriculum designers can foster caption literacy—the ability to use captions strategically for noticing, retrieval, and reflection—within English language programs. By situating captioned YouTube viewing within the lived realities of Bangladeshi students, this study reconceptualises it as a legitimate and valuable form of incidental vocabulary learning that extends the boundaries of formal education into the dynamic, multimodal world of everyday digital engagement.

2. Literature Review

Vocabulary learning has long been recognised as a cornerstone of second language acquisition, yet how learners acquire and retain new lexical items—particularly in informal digital environments—remains an evolving field of inquiry. This section synthesises theoretical and empirical literature across three domains: (a) theories of vocabulary learning, with emphasis on incidental learning, multimodality, and connectivism; (b) global studies on captioned videos and language learning; and (c) research on informal digital learning, particularly within Asian and Bangladeshi contexts. Together, these strands establish the conceptual and empirical grounding for exploring how Bangladeshi university students learn vocabulary through captioned YouTube videos.

2.1 Theoretical Perspectives on Vocabulary Learning

Incidental vocabulary learning refers to the process of acquiring lexical knowledge as a by-product of engaging in meaning-focused communication, rather than through explicit instruction. It is a crucial mechanism for vocabulary expansion in contexts where classroom time and attention are limited. Nation (2013) and Webb (2020) emphasise that while incidental gains from single exposures are often modest, cumulative encounters across multiple contexts lead to durable learning. In multimedia settings, this process depends heavily on noticing—the learner's conscious attention to new lexical items (Schmidt, 2001). Captions, by providing textual reinforcement of spoken words, enhance this noticing mechanism and help consolidate sound–form–meaning connections.

Theories of multimodal learning extend this understanding by examining how learners process input delivered through multiple sensory channels. Mayer's (2020) cognitive theory of multimedia learning and Paivio's (1991) dual coding theory both argue that information is better retained when it is represented both verbally and visually, allowing for richer mental connections. When learners watch captioned videos, auditory, textual, and visual cues converge to support vocabulary recognition and retention. Studies have shown that this dual-channel processing reduces ambiguity and supports deeper lexical encoding, particularly for low-frequency or abstract words. However, multimodality is not inherently beneficial; cognitive load theory warns that excessive or poorly synchronised input can overwhelm working memory (Sweller, 2020). Thus, captions function most effectively when learners can control pace, genre, and caption type, allowing for adaptive engagement.

While incidental learning and multimodality address cognitive mechanisms, connectivism situates learning within the broader digital ecosystem. Siemens (2005) and Downes (2012) propose that in networked environments, knowledge resides not solely in the individual but in the connections formed across platforms, peers, and resources. From this perspective, learning vocabulary through captioned YouTube videos involves navigating a web of interactions—between captions, video content, comment threads, subtitles, and even algorithmic recommendations. Learners actively curate playlists, revisit videos, and cross-reference terms in online dictionaries or social discussions. Such actions demonstrate agency and self-regulated learning that traditional classroom theories often overlook. The integration of incidental learning, multimodality, and connectivism therefore provides a comprehensive framework for understanding how digital learners in Bangladesh experience caption-mediated vocabulary development as a dynamic, networked process rather than a linear cognitive event.

2.2 Studies on Captioned Videos and Vocabulary Learning

Over the past two decades, a substantial body of research has examined the role of captions and subtitles in second language acquisition. Early experimental studies found that same-language captions improved comprehension and vocabulary recall compared to no captions or L1 subtitles (Winke et al., 2010; Vanderplank, 2016). More recent research confirms that captions help learners notice lexical forms, collocations, and phrase boundaries while watching videos (Teng, 2019). Captions enhanced not only word recognition but also listening comprehension, especially for low-frequency words. Similarly, Webb and Rodgers (2009) demonstrated that television viewing with captions exposes learners to a broad lexical range—approximately 3,000–4,000 word families in just 10 hours of viewing—suggesting substantial potential for incidental vocabulary growth.

However, results across studies remain nuanced. Factors such as caption type, proficiency level, and learner autonomy mediate learning outcomes. For instance, Hasan et al. (2019) observed that advanced learners benefitted more from full English captions, while lower-proficiency learners preferred bilingual captions for comprehension scaffolding. Research by Peters and Webb (2018) also showed that learners with larger existing vocabularies are better equipped to notice and retain new words from captioned input, supporting the "Matthew effect" in vocabulary acquisition. Moreover, studies highlight that repeated exposure and learner engagement—rather than mere caption presence—determine retention rates (Vanderplank & Teng, 2024).

Despite these insights, the vast majority of studies have employed quantitative or quasi-experimental designs conducted in controlled classroom environments. Few have explored how learners interact with captioned content in self-directed, everyday digital settings, or how social, affective, and infrastructural factors influence their engagement. Consequently, our understanding of captioned video learning remains incomplete without qualitative perspectives that capture learners' motivations, constraints, and evolving strategies.

2.3 Informal Digital Learning and Social Media Contexts

The rise of informal digital learning has prompted increasing scholarly attention to learner-driven engagement outside institutional settings. Sauro and Zourou (2019) describe such spaces as the "digital wilds," where learners encounter authentic language in naturally occurring online environments. Research on Informal Digital Learning of English (IDLE) suggests that activities such as watching videos, gaming, and interacting on social media can significantly contribute to vocabulary growth, provided learners engage with content actively and reflectively. In East and Southeast Asia, qualitative studies have revealed that students integrate digital resources into everyday routines, often combining entertainment and education through self-selected multimedia content (Lai et al., 2013).

Within South Asia, similar patterns are emerging but remain under-researched. Bangladeshi university students increasingly turn to YouTube and mobile applications for English learning, motivated by convenience, perceived authenticity, and peer influence (Hasan et al., 2019). However, existing studies have focused primarily on general digital learning or pronunciation improvement rather than vocabulary development or caption use. Moreover, infrastructural limitations—intermittent internet connectivity, device constraints, and variable caption quality—shape how learners access and process content. These contextual realities make Bangladesh a compelling site for extending IDLE and multimodal learning research.

2.4 Conceptual Synthesis and Research Gap

Synthesising these domains reveals clear gaps in current scholarship. While cognitive studies confirm that captions can support vocabulary acquisition, they seldom explore the lived experiences of learners who rely on them outside classrooms. Similarly, multimodal learning theories explain processing mechanisms but overlook how learners adapt to technological and linguistic constraints in under-resourced environments. Connectivist perspectives, meanwhile, highlight networked learning

but have rarely been operationalised in empirical language-learning research, especially in the Global South.

Addressing these omissions, the present study advances a qualitative, context-sensitive investigation of how Bangladeshi university students experience and manage vocabulary learning through captioned YouTube videos. It explores how learners perceive affordances and constraints, how they integrate captions into their digital routines, and how these practices extend or challenge traditional notions of vocabulary pedagogy. By situating the study within the interstices of incidental learning, multimodality, and connectivism, this research contributes both theoretically and empirically to a more holistic understanding of informal digital language learning in low-resource educational contexts.

3. Methodology

3.1 Research Design

This study employed a qualitative, phenomenological design to explore how Bangladeshi university students experience vocabulary learning through captioned YouTube videos. The aim was not to measure the extent of learning but to understand the *processes*, *perceptions*, and *strategies* through which learners engage with captioned content in their daily digital practices. A qualitative approach was particularly suited to this inquiry, as it allowed for an in-depth exploration of participants' lived experiences—how they perceive, interpret, and make sense of learning vocabulary beyond classroom boundaries (Milon et al., 2023). This interpretivist orientation aligns with Creswell and Poth's (2018) emphasis on understanding phenomena from participants' perspectives and with Moustakas' (1994) phenomenological concern for meaning-making grounded in individual experience.

In this study, the phenomenological stance was descriptive rather than transcendental, focusing on the *what* and *how* of learners' experiences rather than abstract essence-seeking. Such a stance enabled the researcher to capture rich, contextually embedded narratives of caption use within the sociocultural and technological realities of Bangladeshi higher education. Similar qualitative orientations have been used in recent computer-assisted language learning (CALL) studies that prioritise learner voices and the micro-dynamics of digital engagement (Golonka et al., 2014; Braun & Clarke, 2021). The research was conducted from a constructivist position, assuming that meaning is co-constructed between researcher and participant through dialogue and reflexive interpretation.

3.2 Research Sites and Participants

The study was conducted across three universities in Bangladesh—two private and one public—selected to represent diverse institutional conditions, access to digital technologies, and student demographics. Public universities typically operate under resource constraints and limited internet access, while private institutions tend to be more digitally equipped but vary in pedagogical quality. Selecting both sectors ensured a more holistic understanding of learners' experiences and allowed for comparison across infrastructural and pedagogical contexts.

A total of 15 participants were recruited using purposive sampling (Patton, 2015), as they met specific criteria relevant to the research focus. All participants were undergraduate or postgraduate students who reported using YouTube regularly (at least three times per week) for English learning purposes and had at least six months of experience engaging with captioned videos. Diversity in gender, academic discipline, and proficiency level was sought to capture a range of perspectives. The participants' self-reported English proficiency ranged from intermediate to upper-intermediate, determined through background questionnaires. Each participant was assigned a pseudonym to maintain anonymity, and brief demographic data (e.g., age, field of study, language background, and device use) were collected to contextualise their accounts.

This sampling strategy allowed the study to focus on information-rich cases—learners who could articulate their strategies and reflections on vocabulary learning via captioned content. The participants came from disciplines such as English, Business Administration, Computer Science, and Social Sciences, reflecting the linguistic diversity typical of Bangladeshi tertiary classrooms. By including students from both English and non-English majors, the study captured differences in motivation and strategy related to disciplinary needs and linguistic confidence.

3.3 Data Collection

Data were collected over a three-month period using three complementary methods: semi-structured interviews**,** reflective journals**,** and content logs. These overlapping sources allowed for triangulation and provided a multilayered understanding of learners' experiences (Alam et al., 2024; Denzin, 2017; Milon, 2020; Milon et al., 2017).

Each participant took part in one semi-structured interview lasting between 45 and 60 minutes. Interviews explored participants' habits of watching captioned videos, preferences for caption types (e.g., English, bilingual, or auto-generated), perceived vocabulary learning benefits, and challenges related to technological or cognitive constraints. The interviews were conducted in a bilingual format—English and Bangla—depending on the participant's comfort level, ensuring that nuanced meanings and cultural references were accurately conveyed. Questions were open-ended and iterative, allowing new ideas to emerge as participants narrated their practices and reflections.

In addition to interviews, participants maintained reflective journals for three consecutive weeks. They were asked to document their engagement with captioned YouTube videos, including details such as video type, duration, vocabulary noticed, and how they later used or rehearsed those words. Journals provided valuable insights into ongoing meaning-making processes, capturing immediate reflections that might not surface in retrospective interviews (Ortlipp, 2008).

Finally, content logs were collected from participants who voluntarily recorded the titles and URLs of videos they watched, accompanied by brief notes on specific lexical items encountered and learning strategies employed. These logs were not intended for quantitative analysis but to provide contextual grounding for the qualitative narratives, linking participants' reported strategies to specific video materials. Together, these data sources offered a comprehensive view of both habitual and reflective dimensions of digital vocabulary learning.

3.4 Data Analysis

Data were analysed using Reflexive Thematic Analysis (RTA)**,** following the six-phase process outlined by Braun and Clarke (2019, 2021). This approach was chosen for its flexibility and emphasis on the researcher's active role in meaning-making. The analytic process began with repeated reading of interview transcripts, journals, and content logs to achieve data familiarisation. Initial codes were generated inductively, focusing on recurring features of learners' practices, such as caption toggling, repetition routines, vocabulary recording, and perceived learning outcomes.

Codes were then clustered into potential themes, which were iteratively refined through constant comparison within and across participants. Themes were reviewed for internal coherence and distinctiveness before being defined and named. NVivo 14 software was used for data organisation and coding management, though interpretation remained manual and reflexive. The final thematic structure comprised four major themes: *Captions as Vocabulary Scaffolds*, *Learning Beyond the Classroom*, *Balancing Entertainment and Education*, and *Challenges and Access Issues*.

To enhance trustworthiness**,** several strategies were employed. Member checking was conducted by sharing preliminary interpretations with selected participants, who verified the accuracy and resonance of findings. Peer debriefing with fellow TESOL researchers provided additional analytic scrutiny and helped mitigate researcher bias. Rich, thick description and illustrative quotations were included in reporting to ensure transparency and allow readers to assess transferability (Lincoln & Guba, 1985; Yasmin et al., 2024). Analytical memos were maintained throughout the process to document evolving reflections and theoretical decisions, thereby increasing dependability and reflexive awareness.

3.5 Ethical Considerations

Ethical sensitivity guided every stage of the research process. All participants gave informed consent and were clearly informed about the study's aims, the voluntary nature of participation, and their right to withdraw at any time without consequence. Confidentiality was maintained through pseudonyms and secure digital storage, with no personal identifiers retained in transcripts or reports. To protect privacy and anonymity, all potentially identifying details—such as institutional names, YouTube channels, or video titles—were altered or generalised in presentation.

Recognising the power asymmetry between researcher and participants, particular care was taken to ensure that students did not feel pressured to participate or disclose information beyond their comfort level. Interviews were conducted in informal, participant-selected spaces to foster openness and mutual respect. Ethical responsibility in this study was understood as a relational and continuous process rather than a procedural formality, grounded in attentiveness to participants' well-being, trust, and dignity (Tracy, 2020). This approach aligns with contemporary qualitative ethics, where respect and transparency are viewed as integral to research integrity.

3.6 Reflexive Positionality

The researcher's positionality was central to shaping the study's interpretive process. As a university lecturer in Bangladesh and a TESOL practitioner, the researcher held an insider–outsider duality (Chavez, 2008; Milon et al., 2024). The insider position—shared cultural background, linguistic familiarity, and professional experience—facilitated rapport and empathetic engagement with participants, enabling a deeper understanding of their digital learning practices. Yet, the researcher's institutional role as a teacher also created a degree of distance, requiring constant awareness of how authority could influence participants' openness. To minimise this, interviews were conducted in informal, participant-friendly settings using a mix of Bangla and English to encourage comfort and authenticity.

A reflexive journal was maintained throughout data collection and analysis to record assumptions, emotions, and methodological choices (Alam et al., 2022a, 2022b; Finlay, 2002). The researcher consciously examined how personal experiences with captioned video learning might shape interpretations and sought to counterbalance this through peer debriefing and analytical memoing. Rather than attempting to remove subjectivity, the study embraced reflexivity as a hallmark of qualitative rigor—acknowledging that understanding is co-constructed between researcher and participants and that transparency about the researcher's stance strengthens the study's credibility and ethical integrity.

4. Findings

4.1 Overview of Analytic Process and Themes

The analysis followed Braun and Clarke's (2019, 2021) reflexive approach to thematic analysis, proceeding through iterative cycles of familiarisation, inductive coding, and theme development. Early codes captured repeated learner actions such as pausing, replaying, toggling captions, and noting new vocabulary, as well as reflections on perceived gains and constraints. Through constant comparison across interviews, journals, and logs, four interconnected themes were identified: (1) Captions as Adaptive Vocabulary Scaffolds**,** (2) Learning Beyond the Classroom**,** (3) Balancing Enjoyment and Effort**,** and (4) Navigating Challenges and Constraints**.** Collectively, these themes demonstrate how Bangladeshi university students transform captioned YouTube viewing into a dynamic and situated process of vocabulary learning—one shaped by autonomy, enjoyment, resourcefulness, and the sociotechnical conditions of their educational context.

4.2 Theme 1: Captions as Adaptive Vocabulary Scaffolds

Across participants, captions operated as active scaffolds that mediated comprehension and facilitated vocabulary noticing, retention, and experimentation. Learners consistently described captions as "anchors" linking sound and spelling, especially in fast or accented speech. Nasrin (Business, Year 2) reflected, "When I see the word, I can catch the sound and spell it later." Many students described a recurring "pause–replay–note" cycle, in which they paused to replay segments, confirmed the word via captions, and recorded it in notebooks or digital lists. This iterative process transformed passive viewing into deliberate lexical noticing and reinforced the form–meaning connection over time.

Beyond single words, captions supported collocation and phrase learning. Tuli (MA English) observed, "I realised phrases like 'bear in mind' or 'play a role'—they appear together, not separately." Participants reported that seeing entire lexical bundles in text made discourse markers, prepositional phrases, and academic expressions more visible, thereby heightening their sensitivity to register and formulaic sequences. Some participants noted that the repetition of expressions across multiple videos strengthened retention and facilitated later recall during writing or speech.

Learners also exercised considerable autonomy in how they used captions. Many began with full English captions before gradually reducing reliance as comprehension improved. Mahin (EEE, Year 1) explained, "First I watch with captions, second time without—if I still understand, I know the words stayed." Others alternated caption types according to difficulty level: English-only for familiar content and bilingual captions for dense or technical material. This adaptive use reflected strategic management of cognitive load—students calibrating caption support to task demands rather than following a uniform pattern. However, a few advanced learners reported that captions occasionally distracted them from listening or introduced inaccuracies in auto-generated text. These outlier perspectives highlight that captions were not universally beneficial but functioned as adjustable scaffolds whose utility depended on learner control and genre context.

4.3 Theme 2: Learning Beyond the Classroom

Participants integrated captioned viewing into their everyday lives, transforming informal video consumption into sustained language practice. Many described developing micro-learning routines that leveraged short, repeated exposure over time. Sohana (Sociology, Year 4) explained her "bus ritual": "Every morning I watch one 10-minute video with captions and add three new words to my list." Such small but consistent engagement represented an incremental approach to vocabulary acquisition embedded within students' daily rhythms rather than confined to formal study sessions.

Learners also curated personal lexical ecosystems by manipulating YouTube's digital affordances—using playlists, subscriptions, and transcript searches to target disciplinary or professional vocabulary. Hasan (Economics MSc) maintained separate playlists for "policy talk" and "tech updates," noting that recurring items like *mitigate* or *robust evidence* signalled high-value vocabulary worth recording. By intentionally returning to segments where target words appeared, learners demonstrated not just passive exposure but active lexical tracking. Importantly, many reported transferring caption-derived vocabulary into academic writing and oral communication. Nusrat (MBA) remarked, "After seeing 'trade-off' many times in captions, I finally used it in a class presentation—it felt natural."

Peer collaboration further extended this learning beyond individual effort. Groups of students organised informal "caption challenges," sharing short clips and competing to use new expressions in conversation or written tasks. These collaborative practices transformed isolated learning into socially mediated motivation and accountability. However, some participants acknowledged that their focus sometimes drifted toward entertainment, leaving captions "on by habit" without active engagement. Others limited advanced vocabulary use in assignments to avoid sounding "too Western" or pretentious in local academic settings. Such tensions underscore how informal learning is negotiated within sociocultural norms and self-perceptions of linguistic appropriateness.

4.4 Theme 3: Balancing Enjoyment and Effort

Sustaining long-term engagement required balancing cognitive effort with enjoyment. Learners repeatedly emphasised that captions made challenging videos more comprehensible and thus more enjoyable, helping them persist with content beyond their immediate proficiency level. Arif (CSE, Year 3) explained, "If I can follow the captions, I don't feel lost—I want to keep watching." Participants distinguished between high-yield academic content (e.g., TED Talks, explainers) and lighter entertainment genres (e.g., vlogs, comedy interviews). Many employed a "sandwich method," alternating demanding videos with relaxing ones to maintain interest and prevent fatigue.

Managing cognitive load emerged as a consistent theme. Students adopted strategies such as reducing playback speed for dense lectures, viewing shorter clips in multiple passes, or switching between listening-only and captioned modes. These adaptive techniques illustrate an intuitive grasp of self-regulated multimodal learning, where learners control the timing, pace, and intensity of engagement to avoid overload. Additionally, affective regulation played a vital role in persistence. Priya (Law) noted, "When study feels heavy, I watch something funny—but still in English, so I stay connected." For many, maintaining contact with English—no matter the genre—was more important than the immediate lexical return, reflecting a commitment to continuity over perfection.

4.5 Theme 4: Navigating Challenges and Constraints

Access, accuracy, and linguistic legitimacy emerged as defining contextual challenges. Many learners depended exclusively on smartphones, where small screens, subtitle overlap, and inconsistent connectivity disrupted comprehension. Buffering often desynchronised captions, making simultaneous reading and listening difficult. To manage these issues, students devised workarounds such as downloading videos overnight on campus Wi-Fi or enabling offline viewing during commutes. Such adaptations not only mitigated infrastructural limitations but also introduced predictable learning slots within constrained conditions.

Caption accuracy presented another major obstacle. Auto-generated captions frequently misrepresented specialised terminology or accented speech, leading to confusion. Participants recounted errors like "photo-volcanic" for "photovoltaic" and "marginal pregnancy" for "marginal propensity." In response, learners developed critical media literacy—checking comment sections, cross-verifying transcripts, or manually correcting their notes. This behaviour reflects the emergence of digital discernment skills**,** where learners evaluate rather than passively consume digital input.

Cognitive overload was another recurrent issue. Faced with dense vocabulary, learners prioritised retention efficiency through "lexical triage," focusing on words that recurred across videos or aligned with current academic needs. Mehnaz (Public Administration) summarised, "If it's not useful for my essay, I skip it." This selective strategy balanced motivation with manageability, preventing cognitive fatigue. Finally, sociolinguistic self-awareness influenced how learners appropriated new vocabulary. Some avoided advanced lexical items in local contexts to conform to perceived norms of linguistic modesty. Salman (History) reflected, "When I used 'counterfactual,' my teacher said it sounded foreign—so I changed it." Such instances illustrate the broader negotiation of linguistic identity, where learning English intersects with cultural belonging and audience sensitivity.

Despite these constraints, participants demonstrated remarkable resilience and adaptability. They used bilingual captions strategically, created personal heuristics for vocabulary selection, and adjusted video modes to optimise comprehension. Their resourcefulness highlights that in low-resource environments, effective learning relies not merely on technology but on learners' ability to innovate within limitation.

4.6 Summary of Findings

Overall, captioned YouTube videos supported vocabulary development by facilitating focused noticing, reinforcing contextual meaning, and encouraging productive use in writing and speech. Learners extended their engagement beyond classrooms through micro-learning routines, social collaboration, and identity-driven motivation. Sustained progress depended on balancing enjoyment with cognitive effort, self-regulating engagement, and maintaining affective investment in learning. Yet infrastructural and sociocultural constraints—limited bandwidth, caption inaccuracies, and linguistic expectations—mediated outcomes, prompting creative adaptation and critical engagement. Collectively, the findings depict vocabulary learning as a networked, multimodal, and agentive process, in which captions function as dynamic scaffolds that learners continuously adjust to align with their cognitive, social, and contextual realities.

5. Discussion

This study explored how Bangladeshi university students experience vocabulary learning through captioned YouTube videos, with particular attention to their strategies, motivations, and contextual realities. Guided by four interrelated research questions, the analysis examined (1) how learners engage with captioned videos, (2) which caption types and genres they find supportive, (3) what benefits and challenges they encounter, and (4) how newly learned vocabulary transfers into academic or communicative use. The findings collectively reveal that captioned YouTube viewing operates as an adaptive, agentive, and socially mediated form of informal learning, in which learners orchestrate multimodal input, digital tools, and affective regulation to extend vocabulary development beyond classroom boundaries.

Addressing the first question—how students engage with captioned videos—the study found that learners employed captions not as fixed aids but as dynamic scaffolds that they actively controlled. They frequently adjusted caption visibility, pacing, and replay sequences to focus attention on specific lexical items and collocations. Such behaviour illustrates Schmidt's (2001) *noticing hypothesis* and aligns with Hasan et al.'s (2019) *involvement load* framework, underscoring that attention and engagement, rather than exposure alone, drive incidental vocabulary acquisition. Learners' manipulation of caption modes also reflects Mayer's (2020) cognitive theory of multimedia learning, which posits that meaningful learning occurs when learners manage multimodal input to optimise cognitive resources. In doing so, this study extends earlier quantitative research (e.g., Teng, 2019; Perez et al., 2013) by illustrating the *micro-processes* through which captions foster noticing, reflection, and personalised pacing in authentic digital contexts.

In relation to the second question—preferred caption types and genres—students demonstrated sophisticated awareness of cognitive demand. They used bilingual captions to comprehend dense

academic discourse and English-only captions to focus on collocations and register-appropriate expressions. Learners also strategically alternated between academic explainers and lighter entertainment genres to sustain motivation, demonstrating that pleasure and productivity can coexist in informal learning. This pattern exemplifies the principles of connectivism (Siemens, 2005), wherein learning arises through navigation among interconnected digital nodes such as videos, transcripts, and peer exchanges. Rather than distinguishing formal from informal learning, connectivism conceptualises learning as an ongoing process of *networked exploration*. Students' playlist curation and transcript searches illustrate lexical agency—a hallmark of digital literacy—and reveal that vocabulary learning in online spaces is both iterative and self-organising.

The third research question—learners' perceived benefits and challenges—showed that captioned viewing produced cognitive, affective, and identity-related outcomes. Cognitively, captions promoted noticing of form–meaning relationships and reinforced collocational awareness. Affectively, they reduced anxiety and increased persistence, supporting Ushioda's (2011) view of motivation as an emotional–cognitive negotiation. Socially, learners used linguistic and stylistic cues from presenters to construct professional identities such as "policy analyst" or "software engineer." This finding resonates with Norton's (2013) notion of *investment*, where language learning embodies identity work. At the same time, participants faced tensions between global English norms and local academic expectations, often modifying advanced lexis to align with context. Such negotiation reflects Canagarajah's (2013) theory of *translingual practice*, where learners balance competing ideologies of correctness and appropriateness across audiences.

The challenges—including unreliable internet, caption inaccuracies, and small-screen readability—illustrate the infrastructural inequities shaping digital learning in Bangladesh. Yet participants responded with creative strategies such as offline downloads, triage heuristics, and flexible caption use, demonstrating what Pennycook (2017) describes as *local ingenuity*. These findings challenge deficit perspectives on Global South learners by reframing constraint as a catalyst for innovation. In this context, learner autonomy emerges not from technological abundance but from the capacity to sustain engagement amid limitation. Such adaptive agency extends the current understanding of digital learning by situating self-regulation within structural constraint—a perspective underrepresented in Global North CALL research.

The fourth question—transfer of vocabulary to academic and communicative contexts—revealed that learners consciously integrated caption-derived lexis into essays, presentations, and peer discussions. Repeated encounters across media and social accountability reinforced this transfer: participants revisited videos, used target words in group challenges, and paraphrased captioned phrases during tutorials. These findings affirm Nation's (2013) argument that repetition across contexts consolidates vocabulary knowledge, while also highlighting how incidental learning becomes intentional through reflection and re-use. This cycle of noticing, practice, and social enactment demonstrates that informal digital learning can generate formal linguistic gains when feedback loops connect the two spheres.

The findings further refine multimodal learning theory by problematising the assumption of automatic benefit. Excessive textual input occasionally overloaded learners, confirming Mayer's (2020) cognitive load principle that selective attention among modes enhances learning. Participants' tiered viewing strategy—first for gist, then for vocabulary—represents a self-evolved cognitive design for managing

multimodal input. This insight advances current theory by showing that learners actively *sequence* modalities across time, distributing cognitive effort to sustain comprehension and retention.

Bringing together the three theoretical lenses—incidental learning, multimodality, and connectivism—the study proposes an integrated model of vocabulary development as a networked cognitive–social system. Incidental noticing initiates lexical awareness; multimodal reinforcement consolidates memory; and connectivist navigation sustains exposure through digital networks. This synthesis moves beyond dichotomies of formal/informal and cognitive/social learning, portraying learners as orchestrators who combine technological affordances, social interactions, and multimodal cues to co-construct vocabulary knowledge.

From a reflexive standpoint, the researcher's position as a Bangladeshi university teacher provided insider familiarity with students' linguistic struggles while also allowing analytic distance for critical interpretation. This insider–outsider awareness enhanced sensitivity to contextual realities while preventing over-identification with participants. In line with Braun and Clarke (2019), reflexivity here functioned as a means of transparency, acknowledging how interpretation was shaped through shared cultural background and professional proximity.

Overall, this study situates captioned YouTube learning within the wider ecology of informal, affective, and technologically mediated language learning. By connecting cognitive noticing, multimodal interaction, and digital navigation, it offers a contextually grounded account of vocabulary development in the Global South. Learners emerge not as passive consumers but as strategic agents who transform everyday media engagement into purposeful language learning. Ultimately, the findings reaffirm that meaningful vocabulary growth can—and does—occur beyond the classroom when learners are empowered to connect, reflect, and innovate within their digital worlds.

6. Conclusion

This study investigated how Bangladeshi university students learn English vocabulary through captioned YouTube videos, drawing on the intersecting frameworks of incidental learning, multimodality, and connectivism. Using qualitative data from three universities, it revealed that captioned video engagement constitutes a multimodal, self-directed, and socially mediated process through which learners notice, consolidate, and transfer new vocabulary into academic and communicative use. Captions operated not simply as textual aids but as adaptive scaffolds that learners modulated to manage cognitive load, sustain motivation, and connect words to meaningful contexts. Learning unfolded through iterative informal cycles—watching, noting, applying, and revisiting—embedded within students' everyday digital practices.

Theoretically, the study proposes an integrated model of vocabulary learning that bridges cognitive and sociocultural perspectives. Incidental noticing interacts with multimodal reinforcement and digital networking to form a dynamic, networked process of meaning-making shaped by learner agency and contextual realities. The findings illuminate how Bangladeshi students, often navigating infrastructural and institutional constraints, creatively construct hybrid learning pathways that blend entertainment, academic goals, and autonomy. In doing so, this research contributes to Global South perspectives in computer-assisted language learning (CALL), repositioning learners not as passive recipients of online

input but as active agents negotiating access, relevance, and linguistic legitimacy in unequal technological landscapes.

Pedagogically, the results call for a rethinking of vocabulary instruction in higher education. Teachers should treat captioned videos as authentic, pedagogically valuable resources that complement classroom learning. Explicit instruction in caption literacy—the ability to use captions strategically for noticing collocations, pacing comprehension, and evaluating accuracy—can be integrated into language curricula through short workshops or digital literacy modules. Curriculum designers might encourage video-based vocabulary journals in which students track captioned phrases across genres and reflect on their use in academic writing or presentations. Institutions could also curate captioned video repositories tailored to disciplinary English needs (e.g., business communication, STEM vocabulary) and provide teacher training on guided caption use. Collectively, such initiatives bridge formal curricula and learners' informal practices, aligning pedagogy with the realities of digital learning ecologies.

At the policy level, the findings underscore the importance of digital inclusion in national English education strategies. Enhancing campus internet infrastructure, promoting open educational media, and recognising informal learning outcomes within assessment frameworks would sustain and legitimise learning that already occurs in online spaces. These recommendations align with Sustainable Development Goal 4 on Quality Education, which emphasises equitable access to inclusive and lifelong learning opportunities.

This study acknowledges several limitations. Its qualitative scope and modest participant pool prioritised depth over breadth, and the self-selected sample may represent particularly motivated digital learners. Data relied on self-reports and reflective journals, which may not capture unconscious or incidental learning processes. Future research could adopt longitudinal or mixed-methods designs to track vocabulary retention over time and across proficiency levels. Comparative studies in other Global South contexts could test the transferability of the proposed model, while classroom-based experiments might evaluate the pedagogical impact of explicit caption literacy training.

This research enriches understandings of how multimodal, incidental, and networked learning converge in real-world contexts. It demonstrates that Bangladeshi university students harness digital media with creativity and autonomy, turning informal spaces like YouTube into legitimate arenas of language development. Recognising and integrating these learner-driven practices within formal education is essential for fostering inclusive, context-responsive, and sustainable approaches to English language learning in the digital age.

Acknowledgement: The author extends sincere appreciation to the editor and the anonymous reviewers for their insightful feedback, which greatly enhanced the clarity and quality of this paper.

Funding: This research was conducted without any financial support from funding agencies, institutions, or organizations.

Author's Contribution: The author solely conceptualized, designed, conducted, and wrote the manuscript, and takes full responsibility for its content and accuracy.

Conflict of Interest: The author declares no conflict of interest related to this study.

Declaration of Originality: This manuscript has not been submitted to, nor is under consideration by, any other journal or conference.

Data Availability: All data generated or analysed during this study are retained by the author and are available upon reasonable request.

References

Al Nahar, A., Hira, F. K., & Milon, M. R. K. (2024). From policy to practice: Evaluating the implementation of communicative language teaching (CLT) in Bangladeshi primary schools. *ICRRD Journal, 5*(3), 125–138.

Alam, M. R., Ansarey, D., Halim, H. A., Rana, M. M., Milon, M. R. K., & Mitu, R. K. (2022a). Exploring Bangladeshi university students' willingness to communicate (WTC) in English classes through a qualitative study. *Asian-Pacific Journal of Second and Foreign Language Education, 7*(2), 1–17.

Alam, M. R., Islam, M. S., Ansarey, D., Rana, M. M., Milon, M. R. K., Halim, H. A., Jahan, S., & Rashid, A. (2024). Unveiling the professional identity construction of in-service university English language teachers: Evidence from Bangladesh. *Ampersand, 12,* 100178. https://doi.org/10.1016/j.amper.2024.100178

Alam, M. R., Jahan, S., Milon, M. R. K., Ansarey, D., & Faruque, A. S. U. (2021). Accelerating learners' self-confidence level in second language acquisition: A qualitative study. *ICRRD Quality Index Research Journal, 2*(3), 141–153.

Alam, M. R., Milon, M. R. K., & Rahman, M. K., & Hassan, A. (2022b). Technology application in tourism event, education and training for making a nation's image. In A. Hassan (Ed.), *Technology application in tourism fairs, festivals and events in Asia* (pp. 149–163). Singapore: Springer.

Alam, M. R., Milon, M. R. K., & Sharmin, M. (2018). Hindrances to use and prepare the lesson plan of secondary school teachers in Bangladesh and some recommendations. *Australasian Journal of Business, Social Science and Information Technology, 4*(4), 189–197.

Alam, M. R., Sulaiman, M., Bhuiyan, M. M. R., Islam, M. S., Imam, M. H., Hossen, M. S., & Milon, M. R. K. (2025). Online corrective feedback and self-regulated writing: Exploring student perceptions and challenges in higher education. *World Journal of English Language, 15*(6), 139–150.

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health, 11*(4), 589–597.

Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology, 18*(3), 328–352.

Canagarajah, S. (2013). *Translingual practice: Global Englishes and cosmopolitan relations.* Routledge.

Chavez, C. (2008). Conceptualizing from the inside: Advantages, complications, and demands on insider positionality. *The Qualitative Report, 13*(3), 474–494.

Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). SAGE Publications.

Denzin, N. K. (2017). *The research act: A theoretical introduction to sociological methods* (4th ed.). Routledge.

Downes, S. (2012). *Connectivism and connective knowledge: Essays on meaning and learning networks*. National Research Council Canada.

Finlay, L. (2002). Negotiating the swamp: The opportunity and challenge of reflexivity in research practice. *Qualitative Research, 2*(2), 209–230.

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning, 27*(1), 70–105.

Hasan, M. R., Rashid, R. A., Nuby, M. H. M., & Alam, M. R. (2019). Learning English informally through educational Facebook pages. *International Journal of Innovation, Creativity and Change*, *7*(7), 277-290.

Lai, K. W., Khaddage, F., & Knezek, G. (2013). Blending student technology experiences in formal and informal learning. *Journal of computer assisted learning*, *29*(5), 414-425.

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

Mayer, R. E. (2020). *Multimedia learning* (3rd ed.). Cambridge University Press.

Milon, M. R. K. (2016). Challenges of teaching English at the rural primary schools in Bangladesh: Some recommendations. *ELK Asia Pacific Journal of Social Science, 2*(3), 1–17.

Milon, M. R. K. (2020). A close investigation of the interaction between the individual and the society in *The American Scholar*. *Port City International University Journal, 7*(1–2), 71–79.

Milon, M. R. K., & Ali, T. M. (2023). From language movement to language policy: A critical examination of English in Bangladeshi tertiary education. *ICRRD Journal, 4*(4), 101–115.

Milon, M. R. K., Alam, M. R., & Hossain, M. R. (2018a). A comparative study on the methods and practices of English language teaching in Bangla and English medium schools in Bangladesh. *Australasian Journal of Business, Social Science and Information Technology, 4*(3), 118–126.

Milon, M. R. K., Hossain, M. R., & Alam, M. R. (2018b). Factors influencing dropouts at undergraduate level in private universities of Bangladesh: A case study. *Australasian Journal of Business, Social Science and Information Technology, 4*(4), 177–188.

Milon, M. R. K., Hossain, M. R., & Begum, R. (2017). Women's revolution against the male-dominated society in R. K. Narayan's novels. *ELK Asia Pacific Journal of Social Science, 3*(2), 1–14.

Milon, M. R. K., Imam, M. H., & Muhury, P. (2024). Transforming the landscape of higher education in Bangladesh: Teachers' perspectives on implementing outcome-based education (OBE). *ICRRD Journal, 5*(2), 117–135.

Milon, M. R. K., Ishtiaq, M., Ali, T. M., & Imam, M. S. (2023). Unlocking fluency: Task-based language teaching (TBLT) in tertiary speaking classes—Insights from Bangladeshi teachers and students. *ICRRD Journal, 4*(4), 218–230.

Moustakas, C. (1994). *Phenomenological research methods.* Sage Publications.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Norton, B. (2013). *Identity and language learning: Extending the conversation* (2nd ed.). Multilingual Matters.

Ortlipp, M. (2008). Keeping and using reflective journals in the qualitative research process. *The Qualitative Report, 13*(4), 695–705.

Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford University Press.

Patton, M. Q. (2015). *Qualitative research and evaluation methods* (4th ed.). Sage Publications.

Pennycook, A. (2017). *The cultural politics of English as an international language* (2nd ed.). Routledge.

Perez, M. M., Van Den Noortgate, W., & Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System*, *41*(3), 720-739.

Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition, 40*(3), 551–577.

Sauro, S., & Zourou, K. (2019). What are the digital wilds? *Language Learning & Technology, 23*(1), 1–7.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge University Press.

Siemens, G. (2005). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology and Distance Learning, 2*(1), 3–10.

Sweller, J. (2020). Cognitive load theory and educational technology. *Educational technology research and development*, *68*(1), 1-16.

Teng, F. (2019). Incidental vocabulary learning for primary school students: the effects of L2 caption type and word exposure frequency. *The Australian Educational Researcher*, *46*(1), 113-136.

Tracy, S. J. (2020). *Qualitative research methods: Collecting evidence, crafting analysis, communicating impact* (2nd ed.). Wiley-Blackwell.

Ushioda, E. (2011). Motivating learners to speak as themselves. In G. Murray, X. Gao, & T. Lamb (Eds.), *Identity, motivation and autonomy in language learning* (pp. 11–24). Multilingual Matters.

Vanderplank, R. (2016). *Captioned media in foreign language learning and teaching: Subtitles for the deaf and hard-of-hearing as tools for language learning*. Palgrave Macmillan.

Vanderplank, R., & Teng, M. F. (2024). Intentional vocabulary learning through captioned viewing: Comparing Vanderplank's 'cognitive-affective model'with Gesa and Miralpeix. In *Theory and practice in vocabulary research in digital environments* (pp. 15-38). Routledge.

Webb, S. (2020). *The Routledge handbook of vocabulary studies.* Routledge.

Webb, S., & Rodgers, M. P. (2009). Vocabulary demands of television programs. *Language Learning*, *59*(2), 335-366.

Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology, 14*(1), 65–86.

Yasmin, A., Milon, M. R. K., & Imam, M. H. (2024). An examination of practices and perspectives of task-based language teaching (TBLT) in tertiary literary classes: Insights from Bangladesh. *ICRRD Journal, 5*(2), 101–116.