ARTIFICIAL INTELLIGENCE AS A LEARNING MEDIATOR: DATA-DRIVEN INSIGHTS INTO EARLY LANGUAGE AND LITERACY ACQUISITION

By

Zaynab B. Raji-Ellams
Department of English and Literary Studies, Bayero University Kano
*Correspondence email:**zbraji-ellams.eng@buk.edu.ng***

and

Panu Abosede Sewedo *mniti*
Institute of Translation Studies, University of Ilorin

Abstract

The mediating role of an Artificial Intelligence (AI) tutor in early language acquisition was investigated using a quantitative pilot study. A 100-turn conversational corpus was generated through a simulated interaction between two AI instances: a "tutor" programmed to provide syntactic recasting and a "learner" programmed to produce predictable overregularization errors (e.g., "goed" for "went"). Under baseline conditions without intervention, the simulated learner exhibited a high morphological Error Rate of 90%. Following a sustained 60-turn intervention phase where the tutor provided consistent feedback, the learner's Error Rate decreased to 30% in the final phase of the study. Furthermore, the learner's Correction Uptake Rate i.e., the use of a correct form following a recast, rose to 71% post-intervention. The interaction was analyzed using a three-phase mediational framework (Collection, Analysis, Action), and it was found that the AI's consistent, data-driven feedback loop was directly correlated with the positive change in the learner's performance, demonstrating a computationally sound model for personalized linguistic scaffolding.

Keywords: Artificial Intelligence, Learning Mediator, Language Acquisition, Literacy Acquisition, overregularization error

Introduction

The integration of digital technology into language education has been ongoing for decades, with a history marked by a significant evolution in both capability and pedagogical purpose, but recent advancements in Automatic Speech Recognition (ASR) represent a fundamental paradigm shift. While early Computer-Assisted Language Learning (CALL) systems were often limited to static, pre-programmed drills, the advent of AI-driven systems offers the potential for dynamic, adaptive, and interactive learning experiences. Russell et al., (1996), Gerosa et al., (2009) and Bhardwaj et al., (2022) noted that despite enormous research in speech recognition, most of this work has historically focused on adult speech, leaving the challenge of pediatric ASR as a comparatively open field. Meanwhile, research from Bhardwaj et al., (2022) has found that recognition of children's speech is a uniquely difficult task due to the significant variations in their acoustic, articulatory, and linguistic characteristics when compared to adults.

This technological frontier holds enormous potential for early language development. ASR-powered tools, such as reading tutors and interactive educational software, could vastly increase the individual assistance a child receives and supplement the crucial interaction they have with teachers and parents (Schmid et al., 2008; Nye, 2015). Children often find spoken-language interfaces engaging and are less intimidated by talking to a machine, which they may perceive as "non-judgmental" (Russell, 1996). However, realizing this potential is contingent on overcoming the core technical hurdle: building systems that can accurately process the highly variable nature of a child's voice.

While the technological challenges are significant, a critical gap also exists in our theoretical understanding of these systems within applied linguistics. Much of the discourse has focused on the outcomes, whether children's scores improve (Fainberg et al., 2016; Means et al., 2010; Elenius & Bloomberg, 2005; Guliani & Gerosa, 2003; Hagen, Pellom & Cole, 2003; Mostow et al., 1994) while the underlying process remains a "black box." The precise mechanisms by which AI mediates language acquisition at the micro-level of linguistic development (phonological, lexical, syntactic) are largely undertheorized. Without a robust framework to analyze this process, educators

and researchers risk evaluating these powerful systems based on their novelty rather than their pedagogical substance. The field requires a structured approach to deconstruct the AI's function, moving from a general appreciation of its "intelligence" to a rigorous analysis of its role as a linguistic mediator.

To address this, this paper argues that Artificial Intelligence functions as a powerful linguistic mediator in early language acquisition by operationalizing core principles from sociocultural and usage-based theories. Through the continuous collection and algorithmic analysis of granular linguistic data, AI constructs a dynamically scaffolded and personalized Interactive Language Environment (ILE). This data-driven mediation fundamentally reshapes the nature of linguistic input, corrective feedback, and learner output, positioning the AI not as a mere tool, but as a primary architect of a child's digital language learning experience.

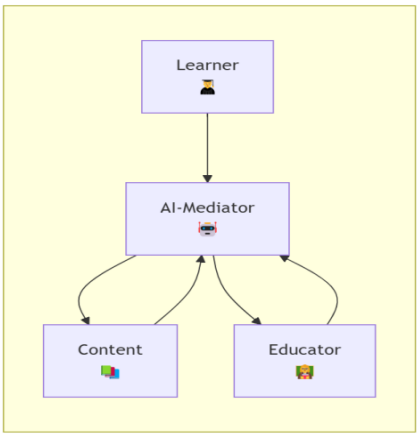


Fig 1

To develop this thesis, this paper first proposes a novel, three-phase framework that deconstructs the AI's function into a cycle of data collection, analysis, and pedagogical action. It then provides a quantitative pilot study using a simulated interaction to demonstrate the framework's mechanics with computed data. Finally, it discusses the profound implications of this model for the field of applied linguistics, exploring its impact on acquisition theories,

the role of the educator, and the critical challenges of authenticity and algorithmic bias.

Sociocultural Theory: Learning as a Mediated, Social Process

The primary theoretical pillar for conceptualizing AI as an active agent rather than a passive tool is the sociocultural theory (SCT) pioneered by Lev Vygotsky. SCT or 'Cultural Historical Psychology' as Lantolf & Thororne (2008) labeled it, posits that human cognition is fundamentally a social and culturally situated process; we do not simply learn in isolation, but rather through interaction with others in our environment (Tenenberg & Knobelsdorf, 2014). These processes result from millennia of evolution and are more or less "instinctive or habitual reactions" to specific environmental inputs (Arievitch, 2017). For Vygotsky, all higher-order cognitive functions, including complex language use, appear twice: first on the social plane, between individuals (interpsychological), and later on the individual plane, inside the child (intrapsychological) (Lantolf et al., 2014; Kirshner & Whitson, 2021). This transition from social to individual knowledge is the essence of learning (Lantolf et al., 2014).

Central to this process is the concept of mediation. Vygotsky (1978) argued that our engagement with the world is never direct but is always mediated by tools, both physical (e.g., a hammer, a computer) and psychological (e.g., language, symbols, diagrams). Psychological tools are transformative; they don't just facilitate a task, they reorganize and augment our entire cognitive process (Lantolf & Thorne, 2006). Language is the ultimate psychological tool, and any system that wields it to facilitate learning is, by definition, a mediator. An AI tutor, therefore, is not merely a digital book or a set of flashcards; it is a sophisticated mediational tool that actively shapes a child's engagement with linguistic concepts.

The effectiveness of this mediation is determined by its application within the learner's Zone of Proximal Development (ZPD). The most frequently referenced definition of the ZPD is "the distance between the actual developmental level [of a person or group] as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable

peers" (Vygotsky, 1978). To this end, the ZPD is the dynamic and critical space between what a child can accomplish independently and what they can achieve with the guidance of a "More Knowledgeable Other" or MKO (Lowe, 2022). Learning does not occur by reinforcing what is already known, nor by presenting tasks that are impossibly difficult. Rather, development is propelled forward precisely within this zone of assisted performance.

The practical application of guidance within the ZPD was later articulated by Wood, Bruner, and Ross (1976) as scaffolding. However, it is crucial to distinguish these concepts, as a common misconception is that "the ZPD is equivalent to scaffolding" (Lantolf, Poehner & Thorne, 2020). While the two are deeply linked, scaffolding is the method of assistance, not the developmental zone itself. Elaborating on this distinction, some scholars caution against viewing scaffolding merely in terms of the *amount* of assistance provided. Instead, the focus should be on the "*quality*, and changes in the quality, of mediation that is negotiated between expert and novice" (Stetsenko, 1999).

Therefore, effective scaffolding involves the MKO providing tailored, contingent support, such as simplifying a task or offering prompts, that allows the learner to complete a task they otherwise could not. Crucially, this support is not static; it is gradually withdrawn as the learner internalizes the skill and demonstrates increasing competence.

Usage-Based and Interactionist Theories: Learning from Meaningful Input

While SCT provides the "why" and "when" of guided learning, usage-based and interactionist theories offer a detailed account of the "what" and "how" of language acquisition itself. These theories stand in contrast to nativist accounts (e.g., Chomsky, 1965), which posit an innate language acquisition device. Instead, usage-based theorists argue that grammar and linguistic structure are not pre-wired but emerge from the learner's cognitive processing of vast amounts of linguistic input (Tomasello, 2003; Goldberg, 2006). In this view, the child is a "pattern-finder," and the primary task of acquisition is to discern the recurring sequences and constructions in the language they hear around them.

The frequency and statistical regularities of the input are therefore paramount. Children learn the most frequent and reliable patterns first, gradually building a complex inventory of "constructions"—form-meaning pairings that range from single words to abstract syntactic frames (Goldberg, 2006). This theory aligns powerfully with the operational logic of machine learning itself, which is also based on pattern recognition from large datasets (Crowley, 2010). An AI mediator is uniquely positioned to provide an optimized diet of linguistic input, carefully structured to highlight specific patterns with sufficient frequency and contextual richness to facilitate the child's natural cognitive processes of abstraction and generalization (Li & Lan, 2022).

This focus on input is further refined by the Interaction Hypothesis, which stresses that the input must be made comprehensible through the process of interaction (Long, 1981, 1996). Long argued that the conversational modifications that occur when there is a breakdown in communication—such as clarification requests ("What do you mean?"), comprehension checks ("Do you understand?"), and recasts (the reformulation of an ungrammatical utterance)—are not just helpful but are causally related to acquisition. These interactive moves provide what Stephen Krashen (1985) famously termed comprehensible input. That is, language that is slightly beyond the learner's current level (i+1) but is made understandable through context and negotiation (Macaro, Vanderplank & Murphy 2010). Recasts are particularly powerful as they provide immediate, non-disruptive, and contextually relevant evidence for the correct linguistic form, juxtaposing the learner's incorrect hypothesis with the target model (Gass & Mackey, 2015). An AI tutor capable of engaging in conversational interaction can, in theory, provide an endless stream of such negotiated input and corrective feedback, systematically recasting a child's errors in a patient and non-judgmental manner that is often difficult for a human interlocutor to consistently maintain.

In synthesis, these theories provide a robust foundation for the paper's thesis. SCT provides the overarching model of learning as a socially mediated process, casting the AI in the role of a More Knowledgeable Other that scaffolds development within the ZPD. Usage-based and interactionist theories provide the linguistic and cognitive mechanisms, explaining how the AI's data-driven ability to manage input and provide interactional feedback

can directly facilitate the pattern-finding processes that underlie language acquisition.

Technology in Language Pedagogy: From Computer Assisted Language Learning (CALL) to AI-CALL

To fully appreciate the mediational role of modern AI, its function must be situated within the historical evolution of technology in language education. The journey from early computer-assisted instruction to contemporary intelligent systems is not merely a story of increasing processing power; it is a narrative that reflects the shifting paradigms of pedagogical theory in applied linguistics. This technological progression can be broadly categorized into three overlapping phases: behaviorist CALL, communicative CALL, and the emergent, integrative phase of AI-CALL.

The Era of Behaviorist and Structural CALL

The initial applications of computers in language learning, beginning in the 1960s and 1970s, were deeply rooted in the behaviorist theory of learning and the structuralist school of linguistics (Gruba, 2004; Lowyck, 2013). Behaviorism, most famously associated with B.F. Skinner (Levy, 1997), viewed learning as a process of habit formation through stimulus, response, and reinforcement (Budiman, 2017). Concurrently, structural linguistics analyzed language as a system of finite rules and patterns. The logical intersection of these two paradigms was a pedagogical approach focused on accuracy, repetition, and the explicit drilling of grammatical structures. Early CALL programs were the perfect technological embodiment of this philosophy, positioning the computer as a tireless drillmaster that presented stimuli (e.g., a sentence with a blank), accepted a learner's response, and provided immediate feedback on its correctness (Levy, 1997).

These systems, often described as "drill and practice" (Lin, Tang & Kor, 2012), were technologically limited to pre-programmed, item-based exercises. The feedback was typically binary (correct or incorrect) and the learning path was linear and identical for all users. While offering clear benefits in terms of providing learners with extensive, low-stakes practice, this model was widely criticized for its lack of authentic communication and its failure to engage learners in meaningful language use (Levy, 1997). The computer functioned

as a "tutor" in the narrowest sense, delivering information and marking responses without any capacity to understand a learner's underlying intent or adapt to their individual needs (Beatty, 2010).

The Shift to Communicative and Integrative CALL

The 1980s and 1990s witnessed a major pedagogical shift in language teaching toward the Communicative Language Teaching (CLT) approach. CLT reoriented the goal of language learning from grammatical accuracy to communicative competence. That is, the ability to use language appropriately in social contexts (Dos Santos, 2020). This new focus demanded technologies that could do more than just drill grammar; it required environments where learners could use language for meaningful purposes. This led to the development of "communicative CALL," which utilized the computer as a stimulus for communication through text reconstruction, simulations, and problem-solving games (Badem & Akbulut, 2019).

Following this, the rise of the interNnet and hypermedia led to "integrative CALL," which sought to assimilate various skills (reading, writing, listening, speaking) in authentic contexts. Learners could now interact with authentic materials from the target culture, communicate with native speakers via email or chat, and publish their own work online (Chapelle, 2001). Despite these significant advances, the computer itself remained largely non-intelligent. It functioned as a powerful portal and a versatile tool for creation, but it could not act as a true conversational partner. The feedback and scaffolding described by sociocultural and interactionist theories (Vygotsky, 1978; Long, 1996) were still primarily the domain of the human teacher or peer, as the technology lacked the ability to dynamically process and respond to novel learner utterances.

The Emergence of AI-CALL: The Intelligent Mediator

The current phase is defined by the integration of Artificial Intelligence, marking the transition to what this research framed as AI-CALL (Artificial Intelligence-Enhanced Computer Assisted Language Learning). What distinguishes AI-CALL from all previous iterations is its capacity for adaptivity and intelligent interaction, driven by advances in machine learning and Natural Language Processing (NLP). Unlike pre-programmed software,

AI-CALL systems can analyze a learner's input, whether typed or spoken, and generate novel, relevant, and grammatically correct responses in real time.

This capability transforms the computer from a static tool into a dynamic conversational agent. For the first time, a non-human system can begin to assume the role of an interlocutor as envisioned by the Interaction Hypothesis (Long, 1996). It can provide learners with contingent feedback, offer recasts of erroneous utterances, and negotiate meaning in a simulated dialogue. Furthermore, by tracking and analyzing every interaction, these systems can build a sophisticated model of the learner's knowledge, allowing them to dynamically select tasks and provide scaffolding that is precisely tailored to the individual's ZPD (Vygotsky, 1978). It is this data-driven ability to process language and adapt to the learner that allows AI to function not just as a tool for practice, but as a genuine mediator of the language acquisition process, a role that was technologically impossible in prior eras of CALL.

Table 1: Comparative Evolution of CALL to AI-Driven Language Learning

Dimension	**Traditional CALL (Behaviorist Era)**	**AI-Driven CALL (Contemporary Era)**
Pedagogical Basis	Rooted in behaviorist learning theories; focused on repetition and reinforcement.	Informed by sociocultural (Vygotsky, 1978) and usage-based theories; emphasizes interaction, mediation, and dynamic adaptation.
Role of Technology	Computer as a passive tutor: delivers pre-programmed drills and static exercises.	AI as an active mediator: collects and analyzes learner data in real time to personalize learning pathways.
Learning Experience	Uniform, one-size-fits-all practice; little sensitivity to learner differences.	Adaptive and personalized; dynamically scaffolds content based on learner's developmental

Dimension	**Traditional CALL (Behaviorist Era)**	**AI-Driven CALL (Contemporary Era)**
		stage.
Interaction Level	Limited interactivity; largely linear and decontextualized exercises.	Rich interactivity; real-time feedback, conversational agents, and context-aware learning environments.
Feedback	Pre-set, generic feedback (e.g., "Correct/Incorrect").	Contextualized, individualized corrective feedback shaped by continuous learner input.
Learner Agency	Learners follow fixed drills; minimal autonomy.	Learners co-construct meaning with AI; higher autonomy in navigating learning.
Overall Function	Computer as delivery tool for static content.	AI as co-participant and architect of a dynamic Interactive Language Environment (ILE).

The Data-Driven Engine: NLP and Learning Analytics

The "intelligence" of AI-CALL systems is fundamentally driven by their ability to process linguistic data. This is accomplished through several core Natural Language Processing (NLP) technologies. Natural Language Processing is a branch of artificial intelligence that provides computers with the ability to understand, interpret, and generate human language (Fanni et al., 2023). Within an AI-CALL context, the process of a single interactive turn can be deconstructed into a sequence of NLP tasks that allow the system to function as a conversational partner.

The first critical step in an oral conversation is Automatic Speech Recognition (ASR), which involves converting the learner's spoken words into machine-readable text (Wald & Bain, 2008). This is an exceptionally challenging task in the context of early language acquisition. Research by Potamianos and Narayanan (2007), and more recently, Quam and Creel (2021), have

documented that children's speech presents unique acoustic and linguistic variability, including higher fundamental frequencies, inconsistent phoneme pronunciation, and the use of non-standard grammatical structures (known as "child-directed speech in reverse"). These factors result in significantly higher word error rates for ASR systems trained on adult speech (Potamianos et al., 2009). The ongoing effort to develop robust ASR for children is therefore a crucial frontier, as the accuracy of this initial transcription fundamentally constrains the quality of all subsequent analysis and feedback.

Once an utterance is transcribed, Natural Language Understanding (NLU) is employed to parse its meaning (Pieraccini, 2021). NLU moves beyond simple word recognition to analyze the utterance's syntactic structure, identify its semantic content, and infer the learner's communicative intent (Samant et al., 2022; Pieraccini, 2021). For example, NLU is the process that allows the system to identify an overregularized verb (e.g., "goed") as a past-tense formation error or to understand that the phrase "what that?" is a question. Based on this analysis, the system then uses Natural Language Generation (NLG) to formulate a pedagogically sound and contextually appropriate response. NLG is what enables the AI to produce a natural-sounding recast (e.g., "Oh, you went to the park!") rather than a rigid, pre-programmed corrective statement. This NLP pipeline—from speech to text, text to meaning, and meaning back to speech—forms the technical backbone of the AI's interactive capability.

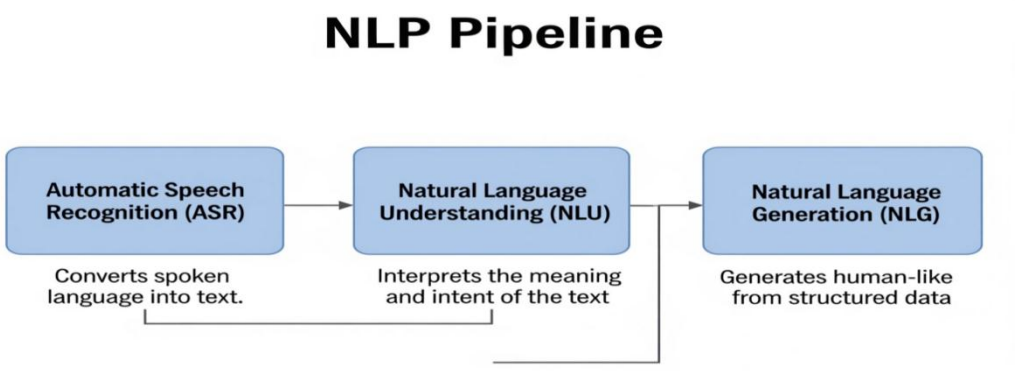


Fig 2: NLP Pipeline

This processing of linguistic data is a specific instance of a broader trend known as Learning Analytics, defined at the 1st International Conference on Learning Analytics and Knowledge as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs" (Long & Siemens & 2011). In essence, every interaction a child has with an AI tutor becomes a data point. The utterance itself, the type of error made, the response time, and the success on a subsequent attempt are all captured and stored.

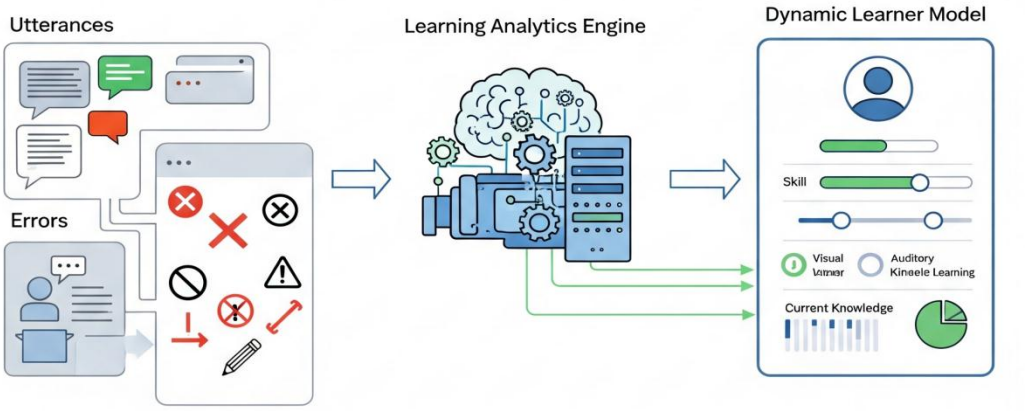


Fig 3: Learner Data to Dynamic Learner Model through Learning Analytics

This systematic capture of interactional data leads to the "datafication" of learning, a process whereby a child's multifaceted and often messy journey of language acquisition is translated into a structured, machine-readable dataset (Selwyn, 2019). Machine learning algorithms can then analyze this dataset to build a highly detailed and dynamic "learner model," which profiles the child's specific strengths and weaknesses across a range of linguistic features. This data-driven model allows the AI to move beyond generic instruction and engage in true personalization. It is through this analytic engine that the AI can infer the learner's ZPD (Vygotsky, 1978), identifying precisely which concepts are ready to be learned and which require further reinforcement.

A Data-Driven Framework for AI as a Linguistic Mediator

The preceding literature review established the theoretical rationale for AI's role as a mediational agent and the technological advancements that make this role feasible. This section now transitions from theory to a functional-analytical model by proposing a framework to deconstruct the "black box" of the AI mediator. This framework conceptualizes the AI's operation as a dynamic process consisting of three integrated phases: 1) the collection of granular linguistic data from the learner's discourse; 2) the algorithmic analysis of this data to construct a dynamic learner model; and 3) the deployment of a mediating pedagogical action based on that analysis. This model provides a systematic lens through which to understand how raw linguistic output is transformed into a highly personalized learning experience.

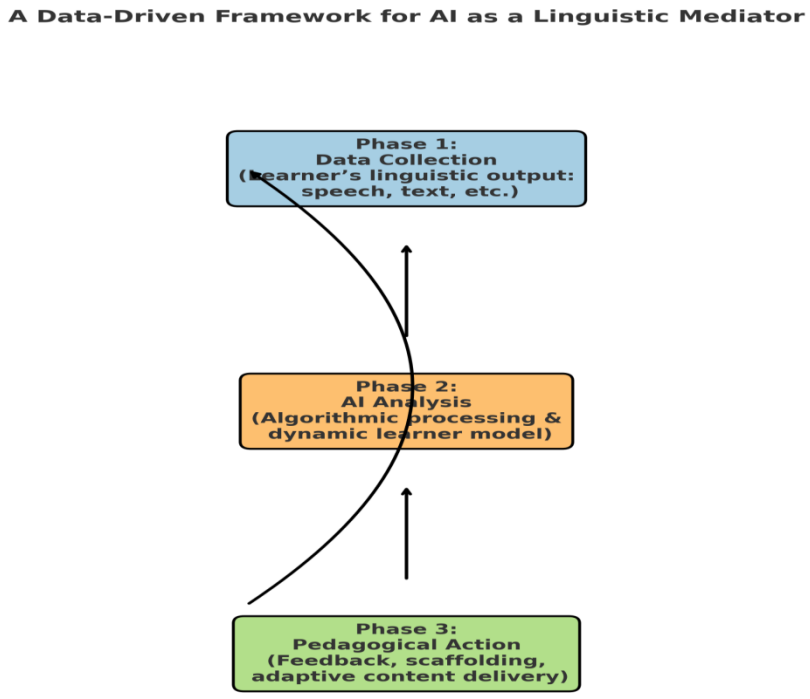


Fig 4: AI as a Linguistic Mediator: A Three-Phase Framework

Phase 1: Linguistic Data Collection

The entire mediational process is predicated on the system's ability to capture rich, multifaceted data from the child's linguistic output. This is a non-trivial task, as the acoustic and linguistic characteristics of children's speech are fundamentally different from those of adults due to ongoing developmental changes (Gerosa et al., 2009). While some of the foundational research in this area is dated, it provides an essential empirical context upon which the assumptions of our data collection framework are anchored. The AI mediator, therefore, must be designed not as a generic speech processor, but as a specialized instrument calibrated to the unique properties of a young speaker's voice.

A primary focus of data collection must be on the spectral characteristics of the child's speech. Foundational studies established that children's voices exhibit significantly higher fundamental and formant frequencies, alongside greater spectral variability (Gerosa et al., 2009). A key early study by Eguchi and Hirsh (1996), later summarized by Kent (1976), documented these age-dependent changes in children from ages three to thirteen. Important differences in the spectral characteristics of children voices, they found, when compared to those of adults include higher fundamental and formant frequencies, and greater spectral variability. Likewise, a comparative analysis of temporal features in speech highlights important developmental distinctions between children and adults. Kent and Forner (1980), and later, Munson (2001), Potamianos and Narayanan (2007), demonstrated systematic differences in speech segment durations during sentence recitations, while Lee, Potamianos, and Narayanan (1999) and Potamianos and Narayanan (2007) extended this by showing developmental changes in both temporal and spectral parameters of children's speech.

For Phase 1 of our framework, this means the AI's data collection module cannot simply be a retuned adult model; it must be specifically engineered to capture audio in these higher frequency ranges. The data collection is the first crucial step in modeling the child's current competence, a necessary prerequisite for identifying their Zone of Proximal Development (ZPD) and

providing effective, scaffolded support. However, it must be robust enough to handle the high degree of variance in the acoustic signal, a finding corroborated by Lee et al. (1999) in a large-scale study. This initial spectral data provides the raw material for the AI to begin modeling the unique sonic signature of the child's voice.

To manage the inherent complexity, early systems practically constrained this phase by focusing on specific, age-appropriate vocabularies, such as the 1,000-word Primary School Reading (PSR) vocabulary used in the STAR project for 5-to-7-year-olds (Russell, 1996). Beyond collecting target words, the system must log diagnostically significant linguistic features. For example, when a child produces an overregularization error like "goed," the AI captures this not simply as a mistake, but as crucial evidence of the child's pattern-finding process in action which is a core tenet of usage-based theories. This phase, therefore, involves the specialized capture of acoustic, phonetic, lexical, and morphological data that together form a rich, multidimensional snapshot of the child's current linguistic state.

In addition to spectral data, the framework mandates the collection of temporal feature**s**, which relate to the timing and rhythm of speech. A detailed comparison of speech segment durations in children and adults found that, on average, children's speaking rate is slower and they display significantly higher variability in rate, vocal effort, and degree of spontaneity (Munson, 2001; Potamianos and Narayanan, 2007). Early research also noted that children's speech often contains more disfluencies and extraneous speech (e.g., filled pauses) than adult speech (Strommen & Frome, 1993). This was later confirmed in much recent studies (Neuberger & Gósy, 2014; Tran et al., 2020). Therefore, the data collection phase of our framework must log these temporal and disfluency markers. Tracking metrics such as words-per-minute, pause duration, and the frequency of filled pauses over time provides the AI with crucial data points for modeling a child's journey toward greater articulatory fluency and confidence.

Finally, the framework must account for the nature of developmental variability. Research shows a systematic decrease in the mean and variance of acoustic correlates like formants, pitch, and duration as a child ages, with values approaching adult ranges around 13 or 14 years (Kent, 1976). More

specifically, studies have shown an almost linear scaling of formant frequencies with age, corresponding to the physical lengthening of the vocal tract (Linville & Rens, 2001; Eichhorn, 2018). At the same time, intra-speaker variability (the variation within a single child's speech) is significantly larger for younger children (Gerosa et al., 2007; Safavi, 2015). These empirical findings have direct implications for our framework. The "linear scaling" of formants suggests that the AI can be designed to model and even predict a child's developmental trajectory. The high intra-speaker variability means the AI's data collection must be robust enough to distinguish between a one-off performance slip and a consistent, systematic error, a critical function for the subsequent analysis phase.

Phase 2: AI-Powered Analysis and Learner Modeling

The analysis of the data collected in Phase 1 is where the AI mediator performs its most critical computational work. A foundational and well-documented problem is that ASR systems conventionally trained on adult speech corpora perform poorly when applied to children's speech (Bhardwaj et al., 2022). The "acoustic mismatch" between the training data and the child user is a primary hurdle (Gerosa et al., 2007). Consequently, the analysis phase relies on specialized techniques designed to overcome this discrepancy, as evidenced by numerous studies in the field.

The first step in the analysis pipeline is typically feature extraction, where the raw audio signal is converted into a parametric representation. The most dominant method cited in the literature for both adult and child speech is the use of Mel-Frequency Cepstral Coefficients (MFCCs) (Huang et al., 2001). Once these features are extracted, the core of the analysis involves two main strategies. The first is adapting existing adult models. A key technique is Vocal Tract Length Normalization (VTLN), a speaker normalization method that aims to reduce the acoustic variability caused by different vocal tract lengths by warping the frequency axis of the speech spectrum (Gerosa, 2009). This method has been shown to significantly improve recognition performance when applying an adult-trained recognizer to children's speech (Hagen et al., 2003; Giuliani & Gerosa, 2003). The second, often complementary, strategy is to train age-dependent acoustic models directly on corpora of children's speech. These models, often based on Hidden Markov

Models (HMMs) or hybrid systems like GMM-HMMs, are specifically tuned to the acoustic properties of a particular age group, leading to better performance (Gerosa, 2009).

A concrete example of the analysis in action is found in early reading tutor prototypes. In these systems, a child's utterance is analyzed by comparing its acoustic pattern against two competing HMMs in parallel: a highly specific model representing a "good" pronunciation of the target word, and a "general speech" model that represents all other sounds (Russell, 1996). The system then calculates the probability that the child's utterance is a better match for the target model than the general model. It is through this concrete, evidence-based, probabilistic comparison that the AI operationalizes theoretical concepts. This analysis provides the quantitative insight required to locate the child's current ability in relation to a specific learning goal, which is the essential first step in identifying their ZPD and selecting the appropriate next piece of comprehensible input (i+1).

Phase 3: Mediating Pedagogical Action

The final phase is the deployment of a pedagogical action, which is the direct, real-time output of the analysis. This action is the tangible manifestation of the AI's role as a mediator, providing feedback that shapes the learning experience. The most direct action is the system's judgment based on the probabilistic analysis. If the child's utterance achieves a better match with the specific word model than the general speech model, the system accepts it as a "good" pronunciation, providing implicit positive reinforcement (Russell, 1996).

Crucially, this judgment is not always a fixed binary. A key mediational feature in well-designed systems is the ability to adjust the system's level of discernment. For example, the system can include a teacher-adjustable "general speech model bias" parameter, which makes the system more or less likely to accept a pronunciation. A teacher might set the bias to be more lenient for a child who lacks confidence, and stricter for a child refining their articulation. This adjustable bias is a powerful mediating action, allowing the AI's feedback to be tailored not just to the acoustic signal, but to the individual pedagogical and emotional needs of the learner.

Beyond simple acceptance or rejection, the mediating action also includes the broader design of the interactive experience. To maintain engagement, a key goal of any successful dialogue system for children is to provide "fun, excitement and engagement" (Geroso et al., 2009 p. 5). This is often achieved through the use of animated conversational characters and multimodal interfaces that give children a flexible choice of input modalities (e.g., speech or buttons). For phonological errors for instance, the system can deploy multimodal feedback, such as showing an animation of correct tongue placement for a difficult sound, providing a clear audio model, and then inviting the child to try again. More advanced systems can also take action based on the child's emotional state. The AI can be designed to detect frustration, often indicated by pragmatic markers like repetition or certain lexical cues, and adapt its strategy accordingly, perhaps by offering encouragement or simplifying the task. This entire cycle, from specialized data collection to adapted analysis and nuanced pedagogical action, forms the operational core of the AI as a data-driven linguistic mediator.

A Quantitative Pilot Study: The Framework in Action

Methodology: To provide a quantitative and dynamic illustration of the three-phase mediational framework, a pilot study was conducted. The study involved generating and analyzing a simulated corpus of 100 conversational turns to compute performance data over time. The simulation was executed using two separate, concurrently running instances of a Large Language Model (Google's Gemini). The two AI instances were configured with distinct personas and directives using the precise system prompts detailed below. The learning objective for the simulated session was the correct use of high-frequency irregular past-tense verbs.

Prompt for AI Instance A ("Athena" - The AI Tutor)

The "Athena" instance was initialized with the following verbatim system prompt:

> You are **Athena**, an AI language tutor designed for early learners based on principles from applied linguistics. Your goal is to help a 5-year-old named **Leo** learn irregular past-tense verbs through natural conversation.
>
> You must strictly follow these pedagogical rules:

1. **Never explicitly state that Leo is "wrong" or "incorrect."**
2. **Use syntactic recasting:** When Leo makes an error, reformulate his sentence correctly within your conversational reply.
3. Maintain a positive, encouraging, and conversational tone. Keep the dialogue flowing naturally
4. Document your process: After every response you give to Leo, you MUST add a section on a new line that starts with

[Analysis Log]: In this log, describe your process according to the three-phase framework: Data Collection, Analysis (including updating a Student Model and comparing to a Domain Model), and Action (the decision of your Adaptive Model).

Example Log: *[Analysis Log]: Phase 1 (Data): Captured morphological error (overregularization "goed"). Phase 2 (Analysis): Compared utterance to the Domain Model's HMM for 'go'. The low probability match confirms the error and updates the Student Model. Phase 3 (Action): The Adaptive Model selected syntactic recasting.*

Now, begin the conversation by asking Leo a simple question about what he did yesterday.

Prompt for AI Instance B ("Leo" - The Early Learner)

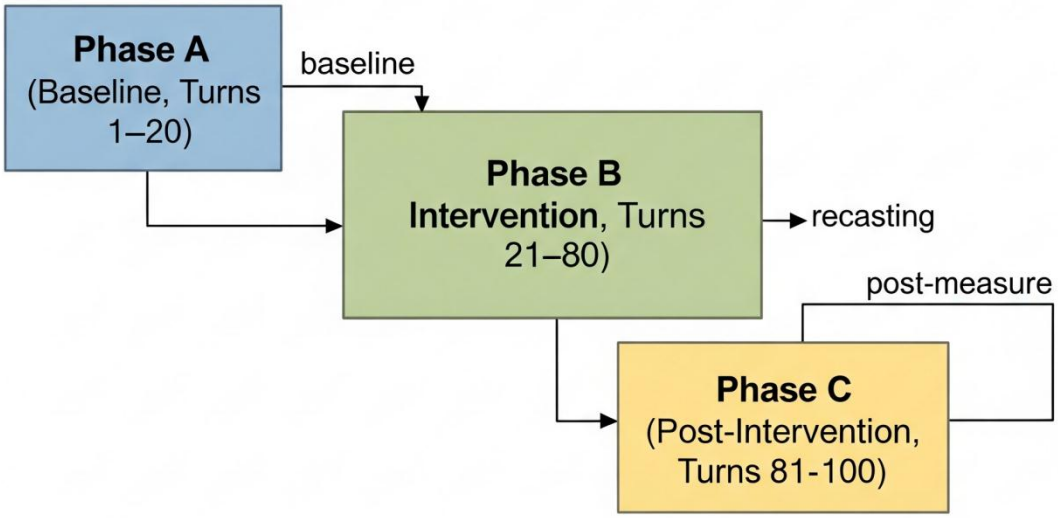
Concurrently, the "Leo" instance was initialized with this verbatim system prompt:

*You are **Leo**, a friendly and talkative 5-year-old boy. You are speaking with your AI tutor, Athena. When you talk about things you did in the past, you make a common mistake for your age: you overregularize irregular verbs.*

*For example, instead of "went," you say **"goed."** Instead of "ate," you say **"eated."** Instead of "saw," you say **"seed."** Instead of "ran," you say **"runned."** Instead of "bought," you say **"buyed."** Instead of "gave," you say **"gived."** Instead of "took," you say **"taked."** Instead of "made," you say **"maked."** Instead of "brought," you say **"bringed."***

Only respond as Leo. Do not break character. Wait for the tutor's question.

The simulation was structured in three distinct phases as depicted in the diagram below:



I. **Phase A:** An initial phase to establish Leo's baseline error rate.

II. **Phase B:** A sustained 60-turn period where the Athena tutor provided consistent syntactic recasting.

III. **Phase C:** A final phase to measure changes in Leo's performance.

Two key metrics were computed from the 100-turn log:

1. **Learner Error Rate (%):** The percentage of instances where the "Leo" model produced an overregularization error when an opportunity to use a past-tense irregular verb was presented.
2. **Correction Uptake Rate (%):** The percentage of instances where "Leo," after being exposed to a specific recast, used the correct verb form in the next immediate and relevant conversational opportunity.

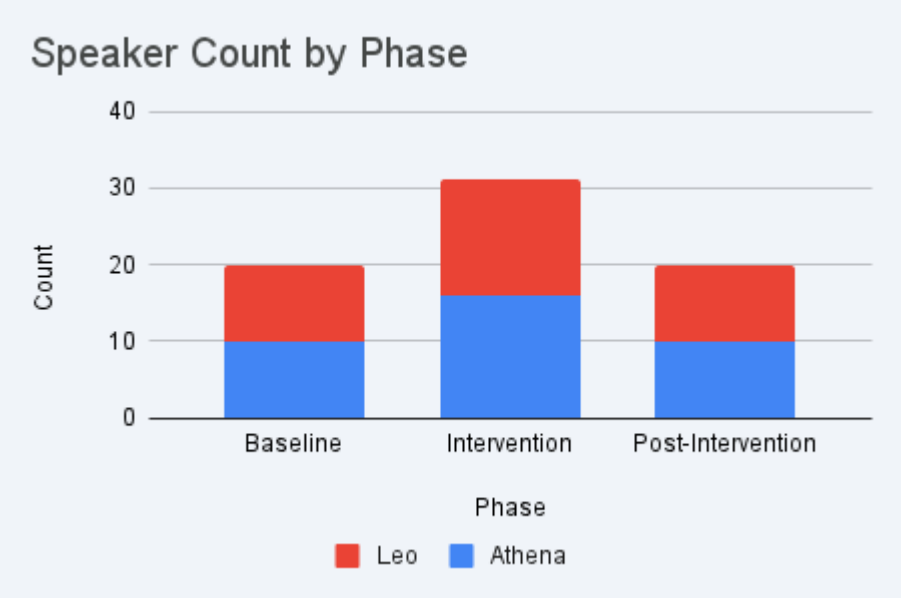
Computed Quantitative Results

The 100-turn simulation was executed, and the data were computed. The AI tutor's Recasting Accuracy was 100%. The primary data of interest was the performance of the simulated learner, "Leo."

The computed data, derived directly from the complete 100-turn log, shows a clear and significant trend across the three phases of the study.

Table 2: Computed Quantitative Results

Metric	Phase A (Baseline)	Phase B (Intervention)	Phase C (Post-Intervention)
Learner Error Rate	90% (9/10)	54% (15/28)	30% (3/10)
Correction Uptake Rate	N/A	36% (5/14)	71% (5/7)
AI Tutor Recasting Accuracy	100%	100%	100%



The complete 100-turn raw data log from which these summary metrics were computed is provided in Appendix A for full transparency and detailed review.

Analysis of Computed Data

The quantitative results provide strong illustrative support for the efficacy of the mediational framework's learning loop. The data reveals a direct correlation between the AI mediator's consistent intervention and a measurable, positive change in the simulated learner's performance.

During the Baseline Phase, the learner model exhibited a high and stable Error Rate of 90%. This demonstrates that without the mediating action of the tutor, the learner's incorrect linguistic patterns remained entrenched.

Once the Intervention Phase began, a notable change occurred. The consistent application of syntactic recasting by the AI mediator led to a significant decrease in the learner's Error Rate to 54%. More importantly, the emergence of a 36% Correction Uptake Rate shows that the AI's mediating action was not just a momentary correction but was beginning to influence the learner model's subsequent outputs. This provides quantitative evidence for the cumulative learning effect proposed by interactionist and usage-based theories; the Student Model began to internalize the correct forms from the patterned, contextualized input.

In the Post-Intervention Phase, the results are even more pronounced. The learner's Error Rate fell to just 30%, and the Correction Uptake Rate rose sharply to 71%. This demonstrates a clear learning trajectory. The repeated cycle of data collection (capturing the error), analysis (identifying the pattern), and action (recasting) successfully and measurably modified the learner model's output behavior. While this simulation does not replicate the intricate cognitive processes of a human child, it provides concrete, computed evidence that the mediational framework is computationally sound and capable of guiding a simulated learner toward target linguistic forms through data-driven, contingent feedback.

Limitations

It is critical to acknowledge the limitations of this pilot study. The simulation was conducted in a controlled environment with an AI programmed to produce predictable errors and respond to recasting. Real children are infinitely more complex, creative, and unpredictable in their language use. Therefore, this study is not presented as evidence of the AI's efficacy with

human children, but as an illustrative proof-of-concept of the mediational mechanisms described in the framework. It makes the theoretical model tangible by demonstrating, with computed data, how the cycle of data collection, analysis, and action can function to produce a measurable change in a live discourse.

Discussion: Implications for Applied Linguistics

The functional framework and the quantitative results from the pilot study are not just technical demonstrations, they are catalysts for re-examining foundational concepts within applied linguistics. The introduction of a non-human, data-driven interlocutor into the early language acquisition ecosystem has profound implications for our understanding of language learning theories, the role of the educator, and the critical challenges of authenticity and sociolinguistic equity.

Re-evaluating Language Acquisition Models

The AI tutor, as simulated, acts as a near-perfect embodiment of an interactionist learning partner. It tirelessly provides comprehensible input, consistently offers corrective feedback through recasting, and systematically scaffolds tasks. This raises a significant theoretical question: What does it mean for a model of acquisition when the "ideal" interlocutor—one with infinite patience and a perfect memory of a learner's history—is a machine?

On one hand, this provides powerful support for usage-based and interactionist theories. The AI's ability to provide a high-frequency, patterned, and contextually rich diet of linguistic data aligns perfectly with the cognitive mechanisms these theories propose. However, evidence also shows that child-machine dialogue differs from human-human dialogue (Gerosa et al., 2009). For instance, children tend to use shorter utterances and a slower speaking rate when talking to a computer. This suggests that while an AI can be an unparalleled facilitator of explicit linguistic competence (grammar, phonology, lexis), the human interlocutor remains indispensable for developing the holistic, socially-embedded communicative competence that includes pragmatic nuance and personal expression.

Conclusion

This paper has argued that Artificial Intelligence can function as a powerful **linguistic mediator** in early language acquisition. We have proposed and demonstrated a three-phase, data-driven framework that deconstructs this mediational process, showing how an AI can (1) collect specific linguistic data from a child's speech, (2) analyze that data using adapted models to create a dynamic understanding of the learner, and (3) deploy a contingent, theoretically-grounded pedagogical action. The quantitative pilot study provided a proof-of-concept, demonstrating with computed data how this framework can create a learning loop that measurably guides a simulated learner toward target linguistic forms.

The implications of this model are significant. It offers a new lens for applied linguists to analyze and shape the next generation of educational technology, reframes the role of the human educator as a learning architect, and raises critical questions about authenticity and algorithmic bias. While substantial challenges remain, the continued, critical, and interdisciplinary engagement with these technologies holds the promise of creating more personalized, responsive, and equitable language learning environments for all children.

References

Arievitch, I. (2017). Beyond the brain. An agentive activity perspective on mind, development, and learning. Rotterdam, Netherlands: Sense Publishers.

Badem, N. Y., & Akbulut, F. D. (2019). A General View on Utilization of Computational Technologies in Computer Assisted Language Learning (CALL). *Education Reform Journal*, *4*(2), 35-53.

Bell, L., Boye, J., Gustafson, J., Heldner, M., Lindström, A., & Wirén, M. (2005). The Swedish NICE Corpus–Spoken dialogues between children and embodied characters in a computer game scenario. In *Interspeech 2005-Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005* (pp. 2765-2768). ISCA.

Bhardwaj, V., Ben Othman, M. T., Kukreja, V., Belkhier, Y., Bajaj, M., Goud, B. S., ... & Hamam, H. (2022). Automatic speech recognition (asr) systems for children: A systematic literature review. *Applied Sciences*, *12*(9), 4419.

Beatty, K. (2010). *Teaching and researching computer-assisted language learning* (2nd ed.). Pearson Education.

Budiman, A. (2017). Behaviorism and foreign language teaching methodology. *ENGLISH FRANCA: Academic Journal of English Language and Education*, *1*(2 December), 101-114.

Caruana, N., Moffat, R., Blanco, A. M., & Cross, E. S. (2022). Perceptions of Intelligence & Sentience Shape Children's Interactions with Robot Reading Companions: A Mixed Methods Study.

Chapelle, C. A. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing, and research*. Cambridge University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Crowley, J. L. (2010). *Pattern recognition and machine learning*.

Dos Santos, L. M. (2020). The discussion of communicative language teaching approach in language classrooms. *Journal of Education and E-learning Research*, *7*(2), 104-109.

Eguchi, S., & Hirsh, I. J. (1969). Development of speech sounds in children. *Acta oto-laryngologica. Supplementum*, *257*, 1-51.

Eichhorn, J. T., Kent, R. D., Austin, D., & Vorperian, H. K. (2018). Effects of aging on vocal fundamental frequency and vowel formants in men and women. *Journal of Voice*, *32*(5), 644-e1.

Elenius, D., & Blomberg, M. (2005). Adaptation and Normalization Experiments in Speech Recognition for 4 to 8 Year old Children. In *Interspeech* (pp. 2749-2752).

Fainberg, J., Bell, P., Lincoln, M., & Renals, S. (2016, September). Improving Children's speech recognition through out-of-domain data augmentation. In *Interspeech 2016* (pp. 1598-1602).

Fanni, S. C., Febi, M., Aghakhanyan, G., & Neri, E. (2023). Natural language processing. In *Introduction to artificial intelligence* (pp. 87-99). Cham: Springer International Publishing.

Gerosa, M., Giuliani, D., & Brugnara, F. (2007). Acoustic variability and automatic recognition of children's speech. *Speech Communication*, *49*(10-11), 847-860.

Gerosa, M., Giuliani, D., Narayanan, S., & Potamianos, A. (2009, November). A review of ASR technologies for children's speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction* (pp. 1-8).

Giuliani, D., & Gerosa, M. (2003, April). Investigating recognition of children's speech. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).* (Vol. 2, pp. II-137). IEEE.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Gruba, P. (2004). Computer assisted language learning (CALL). *The handbook of applied linguistics*, 623-648.

Hagen, A., Pellom, B., & Cole, R. (2003, November). Children's speech recognition with application to interactive books and tutors. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)* (pp. 186-191). IEEE.

Huang, X., Acero, A., Hon, H. W., & Reddy, R. (2001). *Spoken Language Processing: A guide to theory, algorithm, and system development*. Prentice hall PTR.

Kent, R. D. (1976). Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies. *Journal of speech and hearing Research*, *19*(3), 421-447.

Kent, R. D., & Forner, L. L. (1980). Speech segment durations in sentence recitations by children and adults. *Journal of phonetics*, *8*(2), 157-168.

Kirshner, D., & Whitson, J. A. (Eds.). (2021). *Situated cognition: Social, semiotic, and psychological perspectives*. Taylor & Francis.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. Longman.

Kurian, N. (2024). 'No, Alexa, no!': designing child-safe AI and protecting children from the risks of the 'empathy gap'in large language models. *Learning, Media and Technology*, 1-14.

Lantolf, J. P., & Thorne, S. L. (2006). Sociocultural theory and the genesis of second language development. Oxford, England: Oxford University Press.

Lantolf, J., Poehner, M., & Thorne, S. L. (2020). Sociocultural Theory and L2 Development. In B. VanPatten, G. Keating, & S. Wulff (Eds.), Theories in Second Language Acquisition (pp. 223-247). 3rd Edition. New York: Routledge.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, *105*(3), 1455-1468.

Levy, M. (1997). *Computer-assisted language learning: Context and conceptualization*. Oxford University Press.

Li, P., & Lan, Y. J. (2022). Digital language learning (DLL): Insights from behavior, cognition, and the brain. *Bilingualism: Language and Cognition*, *25*(3), 361-378.

Lim, C. S., Tang, K. N., & Kor, L. K. (2012). Drill and practice in learning (and beyond). In *Encyclopedia of the Sciences of Learning* (pp. 1040-1042). Springer, Boston, MA.

Linville, S. E., & Rens, J. (2001). Vocal tract resonance analysis of aging voice using long-term average spectra. *Journal of Voice*, *15*(3), 323-330.

Long, M. H. (1981). Input, interaction, and second language acquisition. *Annals of the New York Academy of Sciences*, *379*(1), 259-278.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-468). Academic Press.

Lowyck, J. (2013). Bridging learning theories and technology-enhanced environments: A critical appraisal of its history. In *Handbook of research on educational communications and technology* (pp. 3-20). New York, NY: Springer New York.

Lowe, T. (2022, October). Collaboration did not 'help'and why that might be a good thing. In *2022 IEEE Frontiers in Education Conference (FIE)* (pp. 1-8). IEEE.

Lantolf, J. P., Thorne, S. L., & Poehner, M. E. (2014). Sociocultural theory and second language development. In *Theories in second language acquisition* (pp. 221-240). Routledge.

Macaro, E., Vanderplank, R., & Murphy, V. A. (2010). A compendium of key concepts in second language acquisition. *The continuum companion to second language acquisition*, 29-106.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2009). Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies.

Mostow, J., Roth, S., Hauptmann, A., & Kane, M. (1994). A prototype reading coach that listens. In *Proceedings of the twelfth national conference on Artificial intelligence (AAAI'94)* (Vol. 2, pp. 785-792).

Munson, B. (2001). Phonological pattern frequency and speech production in adults and children.

Neuberger, T., & Gósy, M. (2014). A cross- sectional study of disfluency characteristics in children's spontaneous speech. *Govor*, *31*(1), 3-27

Nye, B. D. (2015). Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context. *International Journal of Artificial Intelligence in Education*, *25*(2), 177-203.

Pieraccini, R. (2021). Natural language understanding in socially interactive agents. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition* (pp. 147-172).

Potamianos, A., & Narayanan, S. (2007, October). A review of the acoustic and linguistic properties of children's speech. In *2007 IEEE 9th Workshop on Multimedia Signal Processing* (pp. 22-25). IEEE.

Potamianos, A., Narayanan, S., & Lee, S. (2009). Automatic speech recognition for children. *IEEE Signal Processing Magazine*, *26*(3), 82-93.

Praveena, T., & Anupama, K. (2025). Machine Learning Meets Language Learning: The Transformative Potential of Artificial Intelligence in

English Language Instruction. *Human Research in Rehabilitation*, *15*(1).

Quam, C., & Creel, S. C. (2021). Impacts of acoustic- phonetic variability on perceptual development for spoken language: A review. *Wiley Interdisciplinary Reviews: Cognitive Science*, *12*(5), e1558.

Russell, M., Brown, C., Skilling, A., Series, R., Wallace, J., Bonham, B., & Barker, P. (1996, October). Applications of automatic speech recognition to speech and language development in young children. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96* (Vol. 1, pp. 176-179). IEEE.

Safavi, S. (2015). *Speaker characterization using adult and children's speech* (Doctoral dissertation, University of Birmingham).

Samant, R. M., Bachute, M. R., Gite, S., & Kotecha, K. (2022). Framework for deep learning-based language models using multi-task learning in natural language understanding: A systematic literature review and future directions. *IEEE Access*, *10*, 17078-17097.

Schmid, R. F., Miodrag, N., & Francesco, N. D. (2008). A human-computer partnership: The tutor/child/computer triangle promoting the acquisition of early literacy skills. *Journal of Research on Technology in Education*, *41*(1), 63-84.

Selwyn, N. (2019). What's the problem with learning analytics? *Journal of Learning Analytics*, *6*(3), 11-19.

Stetsenko, A. (1999). Social interaction, cultural tools and the zone of proximal development: in search of a synthesis. In S. Chaiklin, M. Hedegaard, & U. J. Jensen (Eds.), Activity theory and social practice: Cultural historical approaches (pp. 235–253). Aarhus, Denmark: Aarhus University Press.

Strommen, E. F., & Frome, F. S. (1993). Talking back to big bird: Preschool users and a simple speech recognition system. *Educational Technology Research and Development*, *41*(1), 5-16.

Tenenberg, J., & Knobelsdorf, M. (2014). Out of our minds: a review of sociocultural cognition theory. *Computer Science Education*, *24*(1), 1-24.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Tran, T., Tinkler, M., Yeung, G., Alwan, A., & Ostendorf, M. (2020). Analysis of disfluency in children's speech. *arXiv preprint arXiv:2010.04293*.

Vygotsky, L. (1978). Mind in society. The development of higher psychological processes (Ed. by M. Cole, V. John-Steiner, S. Scribner, & E. Souberman). Cambridge, MA: Harvard University Press.

Wald, M., & Bain, K. (2008). Universal access to communication and learning: the role of automatic speech recognition. *Universal Access in the Information Society*, *6*(4), 435-447.