Cognitive Foundations of Real-Time Language Communication: Toward a Theoretical Framework of Behavioral Linguistics

Muneo Kitajima
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: mkitajima@kjs.nagaokaut.ac.jp

Makoto Toyota
T-Method
Chiba, Japan
Email: pubmtoyota@mac.com

Jérôme Dinet
Université de Lorraine
Nancy, France
Email: jerome.dinet@univ-lorraine.fr

Katsuko T. Nakahira
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: katsuko@vos.nagaokaut.ac.jp

***Abstract*— This paper explores the cognitive processes involved in real-time language communication during conversations. It emphasizes the dynamic interplay between speakers and listeners, highlighting how both verbal and nonverbal cues contribute to effective communication. The purpose of this paper is to construct a theory of behavioral linguistics on the cognitive architecture, Model Human Processor with Realtime Constraints (MHP/RT), that explains how language is generated in real-time, addressing the limitations of traditional models based solely on Skinner's behavioral psychology for verbal development with System 1 or Chomsky's linguistic theory grounded in the development of formal grammar through System 2, and regarding current natural language being formed through the interaction of Chomsky's grammatical system built upon Skinner's foundation. The proposed theory of behavioral linguistics offers a comprehensive framework for understanding real-time language generation, integrating insights from cognitive psychology and behavioral economics. By framing language use as an action controlled by bounded rationality, this paper highlights the cognitive constraints that influence how individuals communicate. This perspective encourages a more nuanced understanding of conversational competence, suggesting that effective communication relies not only on linguistic knowledge but also on the ability to anticipate and adapt to the dynamic nature of interactions. The implications for improving communication in various contexts, such as education and therapy, underscore the relevance of this research in enhancing everyday interactions and reducing misunderstandings.**

Keywords- Behavioral Linguistics; Everyday Conversation; Verbal Behavior; Nonverbal Communication; MHP/RT, GOMS.

I. INTRODUCTION

For humans, social creatures who use language, everyday conversations with close friends and family bring richness to daily life and enable us to spend fulfilling and happy times. In conversation, the roles of speaker and listener alternate appropriately among participants, maintaining the flow of dialogue. Listeners prepare their own responses while listening to the speaker, anticipating when it will be their turn to speak next in the conversation. The speaker attempts to convey the information they wish to communicate to the listener through both linguistic information transmitted as auditory information and nonverbal information conveyed through visual information such as gestures and eye contact. Skinner [1] conducted a functional analysis of this verbal behavior. Behavior analysts have been working on developing ideas based on verbal behavior for fifty years, and despite this, experience difficulty explaining generative verbal behavior [2]. This suggests that explanations based solely on accumulated conversational information have limitations, and further implies that complex information processing may be involved.

In everyday conversation, speakers adjust their speech to suit the listener's state and deliver appropriate utterances at the right time. Speakers cannot fully grasp the listener's state of understanding, and it is thought that the content of speech is unconsciously selected from several candidates. In verbal behavior, the concepts of bounded rationality and the satisficing principle proposed by Simon [3] are at work. Kahneman proposed the dual-process theory as the cognitive basis for how bounded rationality operates during decision-making [4], laying the foundation for behavioral economics.

This paper proposes that the theory of behavioral linguistics, which can address the verbal behavior of real-time language generation in everyday conversation, can be constructed by examining it on the cognitive foundation of Model Human Processor with Realtime Constraints (MHP/RT) [5]–[7], which concerns action selection under real-time constraints based on an understanding that it is grounded in dual-process theory [4][8]. In MHP/RT, we consider that linguistic behavior emerges through two processes: System-2-Before-Event-Mode, where the listener consciously determines what to say in preparation for becoming the next speaker, and System-1-Before-Event-Mode, where the speaker unconsciously adjusts the content and manner of speech to fit the situation immediately before speaking. Much of verbal behavior is executed unconsciously in a "Goals, Operators, Methods, and Selection rules (GOMS)"-like manner. By employing a rich array of methods concerning mode transitions in the dual-process, it is thought possible to respond instantly to changes in the other party [9]. Furthermore, as a mechanism preventing conversational breakdown, we assume appropriate synchronization—weak synchronization [10]—for these modes

between the speaker and listener. Weak synchronization has been demonstrated as a mechanism for user immersion within virtual environments. In smooth everyday conversation, a state analogous to immersion is thought to manifest. This paper explains these mechanisms that form the foundation of behavioral linguistics.

The main objective of the paper is to propose a theory of behavioral linguistics to explain the real-time generation of language in everyday conversation. This paper is structured as follows. Section II explains the positioning of this research while citing related studies. Section III describes the Perceptual, Cognitive, and Motor (PCM) processes performed by listeners and speakers, and how memory is utilized. Section IV focuses on speaker turn-taking in everyday conversation, elucidating the mechanisms underlying smooth conversational flow. Section V summarizes this research and highlights its relevance to our daily lives.

II. Related Works

The modern scientific analysis of verbal behavior originates in the work of B. F. Skinner [1], whose verbal behavior proposed a functional taxonomy of language based on operant conditioning principles. Skinner's account emphasized environmental contingencies and reinforcement histories as determinants of linguistic behavior. While this framework proved influential, it has been criticized for its difficulties in explaining generativity and the flexible production of novel utterances in spontaneous conversation [11].

To address generativity within a behavior-analytic tradition, Steven C. Hayes and colleagues developed Relational Frame Theory (RFT) [12], which conceptualizes language as generalized relational responding. RFT offers a more powerful explanation of symbolic and rule-governed behavior than classical operant approaches. However, its primary focus lies in derived relational responding and symbolic transformation rather than in the real-time temporal coordination observed in everyday conversational turn-taking.

In cognitive psychology, bounded rationality theory introduced by Herbert A. Simon reframed human decision-making as adaptive under cognitive and environmental constraints [13][14]. Rather than optimizing, individuals select options that are "good enough" given limited time and information. This perspective is particularly relevant to conversational contexts, where speakers must rapidly select utterances without complete knowledge of the listener's internal state.

Building upon bounded rationality, Daniel Kahneman synthesized decades of research on dual-process theory, distinguishing between fast, automatic (System 1) and slow, deliberative (System 2) cognitive processes [15]. Earlier formulations of dual-process models (e.g., [8][16]) emphasized this distinction. While these frameworks have been influential in behavioral economics and decision science, they have not been systematically integrated into a theory of verbal behavior under real-time interactive constraints.

Within psycholinguistics, research on turn-taking demonstrates that listeners begin planning their responses before the current speaker finishes [17]. Conversation analysis has further shown that turn transitions are typically characterized by minimal gaps and overlaps [18]. These findings suggest highly efficient predictive and timing mechanisms, yet they are often treated independently of behavior-analytic theory or broader models of action selection.

In parallel, human performance modeling provides additional relevant foundations. The Model Human Processor (MHP) proposed by Card, Moran, and Newell [19] conceptualizes human activity in terms of Perceptual, Cognitive, and Motor (PCM) subsystems. The Goals, Operators, Methods, and Selection rules (GOMS) framework further formalizes procedural task execution. While these models offer powerful tools for analyzing structured task behavior, they have rarely been extended to spontaneous conversational language generation in naturalistic settings.

So, by integrating all these prior foundations, our paper is innovative for several reasons:

(a) Unlike classical behaviorism [1], the proposed framework incorporates internal action-selection mechanisms grounded in bounded rationality [13] and dual-process cognition [4], enabling an explanation of generative and adaptive utterance formation;
(b) While dual-process theory [15] provides a general cognitive distinction, our paper specifies its implementation within conversational micro-dynamics through two system modes;
(c) By situating verbal behavior within PCM loops [19], our model accounts for the integration of linguistic, nonverbal, and timing behaviors under real-time constraints;
(d) Drawing from findings in conversation analysis [18] and predictive planning research [17], the introduction of weak synchronization [10] provides a mechanistic explanation for smooth turn-taking and breakdown prevention;
(e) Finally, extending traditional MHP and GOMS approaches [19], our proposed MHP/RT [5]–[7] framework explicitly models linguistic action selection under temporal pressure, bridging human performance modeling and behavioral linguistics.

III. Cognitive Process Model of Conversation

A. Description and Scope of the Targeted Conversation Situation

This study focuses on smooth conversation (dialogue) between two individuals. In this conversation, both verbal and nonverbal communication occur. In the former, information is transmitted between participants through auditory information conveyed by sound waves. In the latter, information is transmitted through information from each modality conveyed via the five senses (vision, hearing (including filler information such as interjections), touch, smell, and taste).

This study focuses on verbal communication to delve deeply into the conversation itself. In verbal communication, the observable states of interlocutors can be divided into "speaking" and "not speaking." Therefore, the possible states are as follows:

1) One party is speaking, while the other is not;

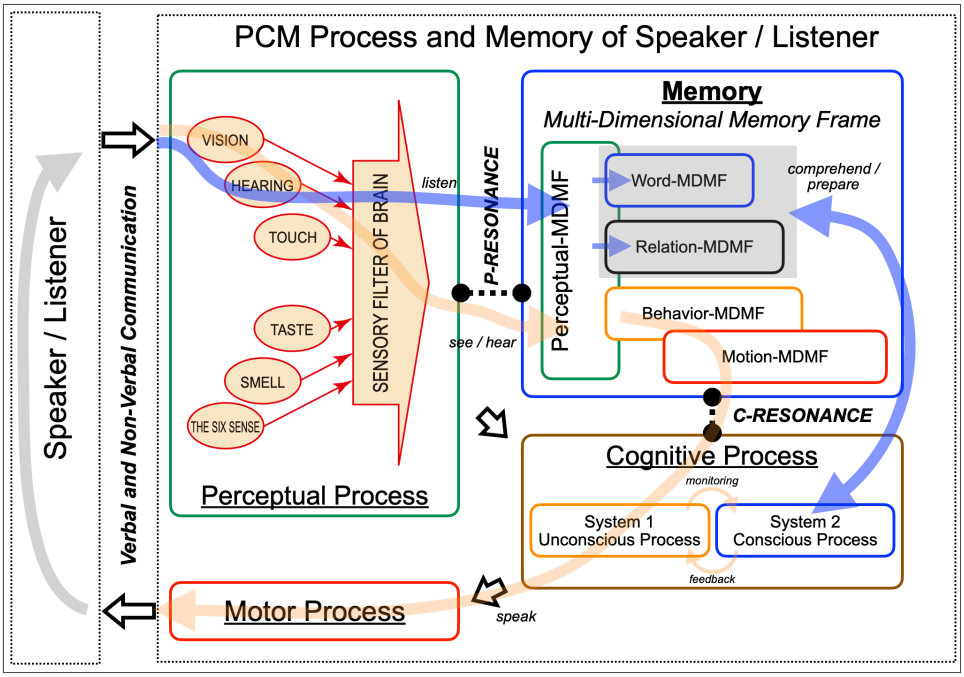

Figure 1. PCM processes and the Multi-Dimensional Memory Frame of conversation participants (modified from [20, Figure 1]).

2) Neither party is speaking;
3) Both parties are speaking simultaneously.

This study examines smooth conversation. In such conversations, both parties concentrate on the conversation and aim to achieve its purpose. Each should be performing verbal behaviors efficiently, without wasting time on unnecessary thought. Therefore, this study assumes the first situation described above. This means that conversation participants have the roles of "speaker" and "listener," and the conversation proceeds through alternating speakers. In this study, the timing of speaker change—that is, the moment when the event of speaker alternation occurs—is considered to be the instant when the speaker changes to the listener.

B. PCM Process Executed by Conversation Participants

We have published cognitive science analyses targeting seamless interaction with the environment [9][21][22]. In these analyses, the environment included Virtual Reality (VR) and other humans. These analyses were grounded in the MHP/RT, a cognitive architecture capable of simulating real-time action selection processes in daily life. This study views everyday conversation as a form of interaction between the environment and humans. By leveraging existing knowledge, we aim to deepen our understanding of the mechanisms enabling seamless conversation.

1) PCM Process and Memory: Conversation participants take on the roles of speaker or listener as the conversation progresses. The speaker actively engages in the conversation, while the listener passively participates. Speaker changes occur at appropriate times, with each person attempting to convey what they wish to communicate to others through verbal and nonverbal communication. When speakers and listeners interact through the format of conversation, sensory nerves within their sensory organs respond to the physical and chemical stimuli emitted by both parties, thereby incorporating information into the individual's brain. Each participant's brain acquires information about its own current activity through multiple sensory organs and generates bodily movements appropriate to the present situation.

Figure 1 illustrates MHP/RT and the process by which information emitted by a conversation participant is incorporated into the body via sensory nerves, undergoes information processing within the brain, and is then expressed as speech via motor nerves, thereby advancing the conversation. This process involves memory, modeled as Multi-Dimensional Memory Frame, and PCM processes. The cognitive process is a dual-process comprising unconscious and conscious processes. This is also referred to as Two Minds, where System 1 executes unconscious processes and System 2 executes conscious processes [4][15]. The Multi-Dimensional Memory Frame consists of Perceptual-, Behavior-, Motor-, Relation-, and Word-Multi-Dimensional Memory Frame. The Perceptual-Multi-Dimensional Memory Frame overlaps with the Behavior-, Relation-, and Word-Multi-Dimensional Memory Frame. This allows activity to propagate from the Perceptual- to Motor-Multi-Dimensional Memory Frame.

Perceptual information taken in from the environment through sensory organs *resonates* with information in the Multi-Dimensional Memory Frame, which is called P-Resonance [20]. In Figure 1, this process is indicated by the symbol •—•. Resonance occurs first in the Perceptual-Multi-Dimensional Memory Frame and activates the memory network. After that,

the activation spreads to the memory networks that overlap with the Perceptual-Multi-Dimensional Memory Frame, and finally to the Motor-Multi-Dimensional Memory Frame.

In cognitive processing based on the Two Minds framework [4][15], conscious processing (System 2) and unconscious processing (System 1) operate in an interrelated manner [20][23]. System 2 utilizes the Word- and Relation-Multi-Dimensional Memory Frame via C-Resonance, while System 1 draws on the Behavior- and Motor-Multi-Dimensional Memory Frame via the same mechanism. Motor sequences are then expressed according to the Motor-Multi-Dimensional Memory Frame. The memories involved in the production of actions are updated to reflect the traces of their use process and influence the future action selection process.

2) PCM Processing and Memory during Conversation: Using Figure 1, we describe processes occurring during conversation: the PCM process and propagation of activation within memory. Regardless of whether a conversation participant is acting as a speaker or listener, information taken into the brain via sensory organs activates the Perceptual-Multi-Dimensional Memory Frame through P-resonance.

The activation propagates within the Multi-Dimensional Memory Frame, undergoing processing via the following two pathways. The first pathway is indicated by the blue arrow in the figure. In this pathway, conscious information processing by System 2 is executed via C-resonance between information propagated from the Perceptual-Multi-Dimensional Memory Frame to Word- and Relation-Multi-Dimensional Memory Frame. The second pathway is indicated by the yellow arrow in the figure. Along this pathway, unconscious information processing by System 1 occurs via C-resonance with information propagated along the Perceptual-, Behavior-, and Motor-Multi-Dimensional Memory Frame. This latter process connects to motor processes, resulting in physical actions reflecting the active state of Motor-Multi-Dimensional Memory Frame.

Speakers and listeners execute the PCM process shown in Figure 1 using the Multi-Dimensional Memory Frame during conversation to perform the following tasks.

The tasks to be performed by the speaker are as follows:

- Utterance: Continuously speaks words based on the selected content;
- Adjustment: Adjusting one's speech content while observing the listener's reaction to one's own utterances or while listening to one's own utterances.

The tasks to be performed by the listener are as follows:

- Understanding: Comprehending the content of the other person's utterances. Here, both verbal and nonverbal information is utilized;
- Nonverbal responses: While listening to the other person's utterances and observing accompanying actions (nonverbal information), one exhibits unconscious nonverbal reactions (eye contact, facial expressions, gestures (nodding), interjections, etc.);
- Preparation: While listening to the other person's utterances, deliberate on what to say after the speaker alternation.

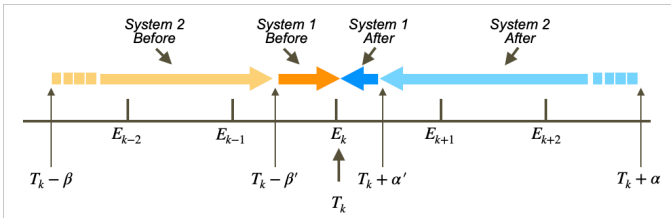

Figure 2. Four processing modes of MHP/RT.

C. Characterization of Conversational Behavior based on the Four Processing Modes

As shown in Figure 1, the PCM process is a cycle. When a speaker hears their own utterance, notices a mispronunciation, and corrects it, the process connects as follows: Motor (utterance) → Perception (listening) → Cognition (detecting the mispronunciation, deciding on a correction method) → Motor (utterance) → $\cdots$. In this way, the PCM process runs as a continuous cycle, but by breaking it at consciously recognized events, it can be perceived as a sequence of events. When considering a conversational behavior, a representative event is speaker alternation. Alternatively, one may become conscious of their own previous utterances while speaking, as if noticing a slip of the tongue. This too is an event occurring during conversation. By focusing on such events, the unbroken PCM cycle can be captured through the following four processing modes.

1) Four Processing Modes of MHP/RT : The experience associated with an individual's activity is characterized by a series of events that are consciously recognized serially. Let $E(T_k)$ denote the event that occurred at time T_k. The experience is then defined as a series of events along the timeline as follows:

$$\cdots \rightarrow E(T_{k-1}) \rightarrow E(T_k) \rightarrow E(T_{k+1}) \rightarrow \cdots$$

Considering the way System 1 and System 2 are involved in individual events, four processing modes can be defined as shown in Figure 2.

Let us focus on an event that occurs at time T_k. For an "event $E(T_k)$" that should occur at time T_k, there exist System 2 conscious processes and System 1 unconscious processes related to $E(T_k)$ before that time T_k. Also, for the "executed event $E(T_k)$" at time T_k, there exist unconscious processes of System 1 and conscious processes of System 2 involving $E(T_k)$ after that time T_k. MHP/RT's System 1 and System 2 operate before and after the event $E(T_k)$ in one of four processing modes for this event.

a) Before the Event ($T < T_k$) : The event $E(T_k)$ that occurs at time T_k reflects the result of the resonance (P-Resonance) between the Multi-Dimensional Memory Frame and the perceptual and cognitive systems—System 1 and System 2—during the time before T_k. $E(T_k)$ is generated by the activities of System 1 and System 2 in the time period before T_k. The different time bands of processing activities result in two processing modes before the event (corresponding

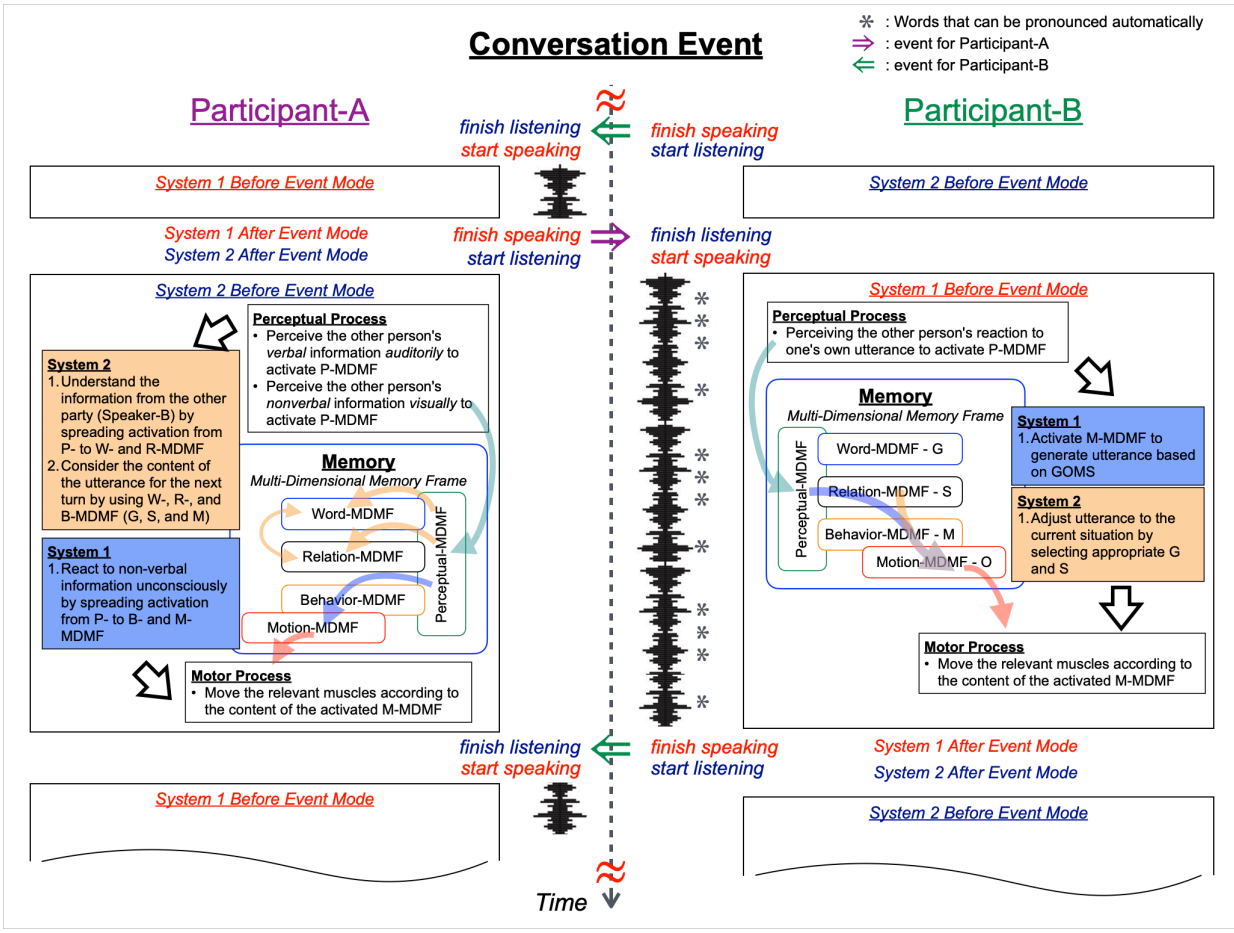

Figure 3. Basic PCM processes for listeners and speakers in conversation and the use of memory.

to the part before time T_k in Figure 2). The two processing modes are:

▷ [System-2-Before-Event-Mode]:
In the time range of $T_k - \beta \leq t < T_k - \beta'$, MHP/RT plans for future events to occur. There is enough time to think carefully.

▷ [System-1-Before-Event-Mode]:
In the time range of $T_k - \beta' \leq t < T_k$, the action selections smoothly generate the immediate event. Here, a series of action selections is executed through feedforward processing led by System 1. During this time, System 2 evaluates the results of the action selections in a timely manner. If it determines that the system is likely to deviate from the expected trajectory or has already deviated, it issues instructions to System 1 for trajectory correction.

Here, $\beta > \beta'$, $150\text{msec} < \beta' < T_k - T_{k-1}$, and β ranges from seconds to hours and months.

b) After the Event ($T > T_k$) : When event $E(T_k)$ occurs at time T_k, the result is stored. Actions occur by integrating the resonances that emerge through interacting with the environment prior to the event, and after the actions are taken, they are bundled and collected. The existing Multi-Dimensional Memory Frame are updated to reflect the results of $E(T_k)$ by the activities of System 1 and System 2 during the time period after T_k. The different time bands of processing activities result in two processing modes after the event (corresponding to the part after time T_k in Figure 2).

▷ [System-1-After-Event-Mode]:
In the time range of $T_k < t \leq T + \alpha'$, to perform better for the same event that may be encountered in the future, the connection between the incoming perceptual information and the output motor content is adjusted unconsciously.

▷ [System-2-After-Event-Mode]:
In the time range of $T_k + \alpha' < t \leq T_k + \alpha$, the event is reviewed and reflected upon. The results are stored and used in the next System-2-Before-Event-Mode before a similar event occurs.

The minimum value of α' is $\sim$ 150msec, and α ranges from seconds to months. In these two modes, action selection results for the event at T_k would be reflected in the network connections of the respective Multi-Dimensional Memory Frame.

2) Basic PCM Process for Conversational Behavior: Figure 3 shows the PCM process and memory during a conversation between conversation participant A and conversation participant B. The speaker executes utterance and adjustment,

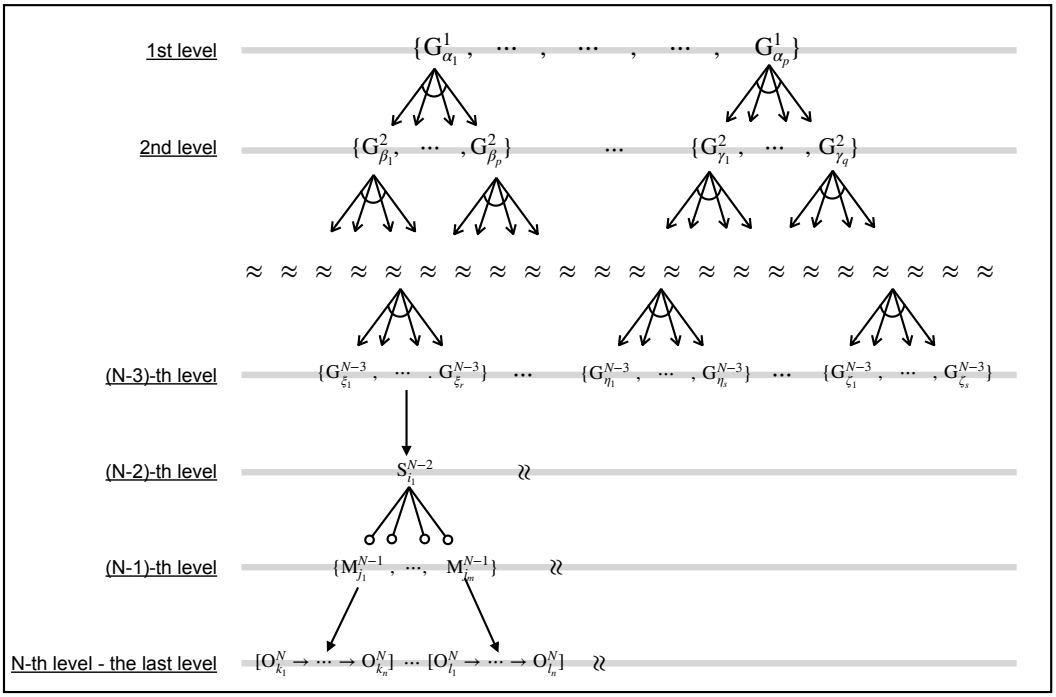

Figure 4. GOMS connection structure [23, Figure 8].

while the listener executes understanding, nonverbal response, and preparation by operating the PCM process (see Section III-B2).

In conversation, cognitive processes are executed through the four processing modes described in Section III-C1. The four processing modes are defined using the time at which an "event" occurs. Since conversations are organized and progress through timely speaker alternations, this study defines the basic event in conversation as "the time when the speaker finishes their utterance," as shown in Figure 3.

The left box in Figure 3 shows the process when conversation participant A is listening to conversation participant B's utterance. The right box in Figure 3 shows the process when conversation participant B is speaking. The processes described within these boxes are explained below.

a) Understanding the Content of Utterance: This process is described in System 2-1 in the left box of Figure 3. Propagation of activation within the Multi-Dimensional Memory Frame is indicated by green and yellow arrows. The speaker's utterance (verbal information) is incorporated as auditory information to activate the Perceptual-Multi-Dimensional Memory Frame. This activation propagates to the Word- and Relation-Multi-Dimensional Memory Frame, enabling understanding of the utterance content in System-2-Before-Event-Mode. The understanding is represented as an activation pattern within the Multi-Dimensional Memory Frame.

b) Preparation for Utterance: This process is described in System 2-2 in the left box of Figure 3. Propagation of activation within the Multi-Dimensional Memory Frame occurs between the Word- and Relation-Multi-Dimensional Memory Frame. This is indicated by the yellow arrows connecting them. Preparation of utterances is performed by System-2-Before-Event-Mode, utilizing the knowledge described below.

Smooth conversation is executed as a routine goal-oriented task. Therefore, this study represents the knowledge utilized during speech using the GOMS model, which consists of goals, operators, methods, and selection rules. Figure 4 illustrates the connection structure of G, O, M, and S. Layers 1 through $(N-3)$ correspond to the goal structure, layer $(N-2)$ to selection rules, layer $(N-1)$ to methods, and layer N to operators [19]. In the GOMS model, the goal structure G is stored in the Word-Multi-Dimensional Memory Frame, and the selection rules S, which determine the appropriate method to apply based on the situation, are stored in the Relation-Multi-Dimensional Memory Frame. These are objects consciously manipulated by System 2. The methods M, which are pointers to operator sequences, are stored in the Behavior-Multi-Dimensional Memory Frame. These are linked to the Relation-Multi-Dimensional Memory Frame. The operators O are stored in the Motor-Multi-Dimensional Memory Frame, and the operator sequences defined by the methods are passed to the motor process, where actions are executed [23].

Applying the general GOMS model to conversational behavior yields the following: the Word-Multi-Dimensional Memory Frame represents the information (goal) the speaker wishes to convey to the listener in utterances following turn-taking. The Behavior-Multi-Dimensional Memory Frame represents the utterance content (method), expressed as the sequence of words stored in the Motor-Multi-Dimensional Memory Frame that is generated automatically and unconsciously. The Relation-Multi-Dimensional Memory Frame represents the candidate selection rules (selection rules) for timely substitution of utterance content based on the listener's situation and the speaker's own situation [19][23].

In the GOMS model used for conversational behavior, the components related to utterance preparation (see Section III-C2b) are as follows. The activation patterns of the Multi-Dimensional Memory Frame generated by utterance understanding (see Section III-C2a) and activated in relation to nonverbal reactions are reflected in the conversation goal structure G within the activated Word-Multi-Dimensional Memory Frame and the selection rules S within the Relation-Multi-Dimensional Memory Frame as the conversation progresses. This process effectively updates the activation state of the GOMS connection structure and places the utterance method candidates stored in the Behavior-Multi-Dimensional Memory Frame to be executed after speaker turn-taking into a standby state.

c) Nonverbal Reaction: This process is described in System 1-1 in the left box of Figure 3. The propagation of activation within the Multi-Dimensional Memory Frame is indicated by green, blue, and red arrows. The listener perceives nonverbal information—facial expressions, gaze, blinking, nodding, touch, murmurs, etc.—emitted by the speaker during utterance through the five senses as perceptual information, thereby activating the Perceptual-Multi-Dimensional Memory Frame. Simultaneously, the Perceptual-Multi-Dimensional Memory Frame overlaps with the Word- and Relation-Multi-Dimensional Memory Frame, which are activated by speech understanding and speech preparation occurring in System-2-Before-Event-Mode. Therefore, Perceptual-Multi-Dimensional Memory Frame activation occurs through these two pathways. The listener exhibits nonverbal behavior via System 1 that reflects the propagation of activation from the Perceptual- to Behavior- and Motor-Multi-Dimensional Memory Frame.

d) Generation of Utterance: This process is described in System 1-1 in the right box of Figure 3. The propagation of activation within the Multi-Dimensional Memory Frame is indicated by the yellow arrows. One candidate suitable for the current situation is selected by applying the selection rules to the speech method candidates in the Behavior-Multi-Dimensional Memory Frame that are placed in a standby state during speech preparation while listening. Then activation is propagated to the Motor-Multi-Dimensional Memory Frame. Following the activation pattern, the motor process is activated to produce speech (red arrow). This process is executed by System-1-Before-Event-Mode. Upon utterance completion, the utterance becomes conscious as an event, and then System-1-After-Event-Mode and System-2-After-Event-Mode are executed. In System-2-After-Event-Mode, the completed goal is deactivated, and processing moves to the next stage: System-2-Before-Event-Mode as listener.

e) Adjustment of the Contents of Utterance: This process is described in System 2-1 in the right box of Figure 3. The propagation of activation within the Multi-Dimensional Memory Frame is indicated by green and blue arrows. The speaker's own utterance is input auditorily and activates the Perceptual-Multi-Dimensional Memory Frame. Furthermore, the listener's nonverbal responses are also perceived through the five senses, activating the Perceptual-Multi-Dimensional Memory Frame. From there, activation propagates to the Word-, Relation-, and Behavior-Multi-Dimensional Memory Frame.

One's own utterances are generated reflecting the activation pattern of the goal structure within the Multi-Dimensional Memory Frame at the time of speaking. The activation pattern within the Multi-Dimensional Memory Frame, triggered by one's own utterance and the listener's nonverbal responses, is processed in a timely manner via System 2 mediated by C-resonance. This process checks whether there is any discrepancy between this pattern and the activation pattern of the goal structure within the Multi-Dimensional Memory Frame that was activated during speech preparation. When discrepancies are detected, the goal and method are reselected using the activation pattern of the goal structure, which is updated sequentially with utterance via System-2-Before-Event-Mode. The utterance is then continued via System-1-Before-Event-Mode (red arrow).

IV. Discussion

A. Analysis of Conversations Involving Speaker Alternation Based on the Four Processing Modes of MHP/RT

Section III-C2 focused on a single speech event and used Figure 3 to explain the basic processes of PCM and memory of the listener and speaker. Actual conversations are executed by linking together these basic processes. Let E_N denote the utterance termination event for conversation participant B in the basic process shown in Figure 3. We will then examine the relationship between E_N and the events that occurred leading up to it.

1) Smooth Conversation Mode: In the smooth conversation mode, the only event that can be consciously recognized is the end of an utterance. Figure 5 illustrates how the conversation progresses, focusing on the four processing modes of MHP/RT. The utterance termination events indicated by "$\Rightarrow$" for the conversation participant A shown on the left are denoted by E_{N-3}, E_{N-1}, while the utterance termination events indicated by "$\Leftarrow$" for the conversation participant B shown on the right are denoted by E_{N-2}, E_N. The utterance is executed by System-1-Before-Event-Mode, and during the utterance, no correction is performed based on the monitoring results from System 2.

We will denote the four processing modes of MHP/RT associated with the utterance termination event E_i as follows.

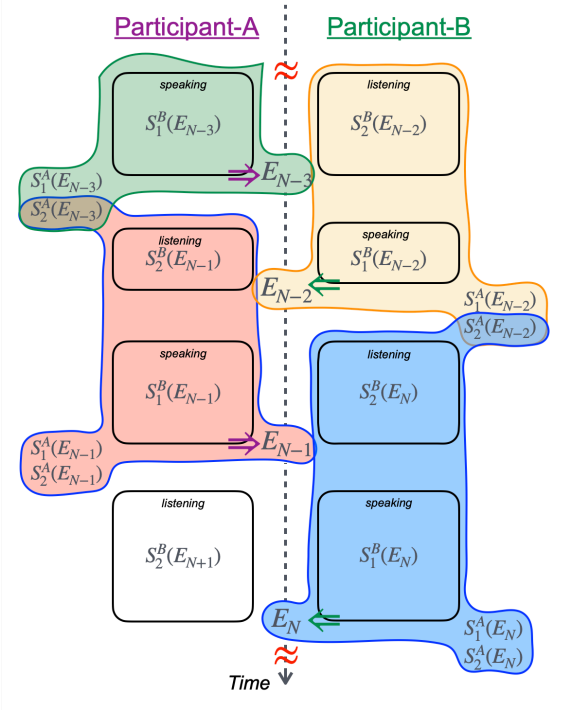

Figure 5. Four processing modes in conversation.

$S_2^B(E_i)$: System-2-Before-Event-Mode
$S_1^B(E_i)$: System-1-Before-Event-Mode
$S_1^A(E_i)$: System-1-After-Event-Mode
$S_2^A(E_i)$: System-2-After-Event-Mode

The parameter β shown in Figure 2 corresponds to the start time of $S_2^B(E_i)$, where $\beta \approx T_i - T_{i-2}$. β' corresponds to the start time of $S_1^B(E_i)$, where $\beta' \approx T_i - T_{i-1}$.

In Figure 5, the portion directly related to participant B's utterance termination event E_N is highlighted with a blue background. Organizing what takes place here in the order of occurrence leads to the following description.

1) While listening to participant A's utterance in $S_1^B(E_{N-1})$, execute one's own utterance preparation in $S_2^B(E_N)$.
2) E_{N-1} occurs when the conversation participant A finishes utterance.
3) Execute one's own utterance in $S_1^B(E_N)$.
4) Finishes one's own utterance, and E_N occurs.
5) Adjust the connection state of Multi-Dimensional Memory Frame nodes unconsciously in $S_1^A(E_N)$.
6) Reflect consciously on E_N in $S_2^A(E_N)$. This will cause changes in the activation patterns in the Multi-Dimensional Memory Frame.

In item 1, the fact that utterance preparation occurs under the conversational partner's utterance is denoted as follows.

$$S_1^B(E_{N-1}) \hookrightarrow S_2^B(E_N)$$

In utterance preparation, the Multi-Dimensional Memory Frame is employed, reflecting the outcome of conscious reflection in $S_2^A(E_{N-2})$ on the event E_{N-2} of one's previous utterance. The Multi-Dimensional Memory Frame of conversation participant B used during utterance preparation reflects the conversation experience E_{N-4}, E_{N-6}, $\cdots$ with conversation participant A up to this point. This is expressed as follows.

$$\cdots,\ S_2^A(E_{N-4}),\ S_2^A(E_{N-2}) \mapsto S_2^B(E_N)$$

The act of generating verbal behavior based on the prepared results is denoted as follows.

$$S_2^B(E_N) \mapsto S_1^B(E_N)$$

To summarize the above, the relationship between the event E_N where the conversation participant B terminates the utterance performed in $S_1^B(E_N)$ and the four processing modes of both participants is as follows.

$$\begin{aligned} &\text{Participant}_\text{A}\colon\ S_1^B(E_{N-1}) \ \hookrightarrow \\ &\text{Participant}_\text{B}\colon\ S_2^A(E_{N-2}) \ \mapsto\ S_2^B(E_N) \mapsto S_1^B(E_N) \end{aligned} \tag{1}$$

The relationship between the immediate preceding event, namely event E_{N-1} where the conversation participant A terminates utterance performed in $S_1^B(E_{N-1})$, and the four processing modes of both participants is as follows.

$$\begin{aligned} &\text{Participant}_\text{A}\colon\ S_2^A(E_{N-3}) \ \mapsto\ S_2^B(E_{N-1}) \mapsto S_1^B(E_{N-1}) \\ &\text{Participant}_\text{B}\colon\ S_1^B(E_{N-2}) \ \hookrightarrow \end{aligned} \tag{2}$$

In the smooth conversation mode, (2) connects to (1), generating E_N. A similar relationship applies backward in time. Thus, it becomes clear that a speaker's utterances are influenced by one's own previous utterances and by other's utterances.

2) Intermittent Conversation Mode: When an utterance is being executed in System-1-Before-Event-Mode, System 2 monitors the utterance content in real time to verify that the method is being executed correctly. If no issues are detected, processing by System 1 continues. At this time, the conversation proceeds in the smooth conversation mode. Conversely, when the utterance is judged to have been executed improperly, the method could be reselected and, if necessary, the goal could be re-established. System 2 interrupts the utterance executed in System-1-Before-Event-Mode and, after the cognitive processing in System-1-After-Event-Mode and System-2-After-Event-Mode, the next utterance is executed in System-2-Before-Event-Mode. In this situation, events occur during the speaker's turn without any speaker change. Therefore, the mode in which conversation proceeds in this manner is called the intermittent conversation mode.

Using Figure 3 to explain, conversation participant B begins utterance in $S_1^B(E_N)$ after the turn-taking event E_{N-1}. However, during the process leading up to the next turn-taking event E_N, an interruption by System 2 occurs, and the situation at that moment is made into an event and becomes conscious. Here, "adjustment" (see Section III-C2e) is performed. The reflection of that interruption event is

performed in System-2-After-Event-Mode, and the prepared actions (see Section III-C2b) are modified to reflect the active state of Multi-Dimensional Memory Frame at that time. This involves modifying the utterance goal, which is performed by referencing the goal structure deployed at the Word- and Relation-Multi-Dimensional Memory Frame levels.

The i-th interruption event occurring during the process leading up to event E_N—where conversation participant B finishes speaking and the turn transitions to conversation participant A, as shown in Figure 3—is denoted as $E_{N,i}$. By the time E_N occurs, the following events have taken place:

$$
\begin{aligned}
&\text{Participant}_{\text{B}}\text{:}\\
&S_2^B(E_{N,1}) \rightarrow S_1^B(E_{N,1}) \rightarrow \quad E_{N,1}\\
&\qquad\qquad\qquad\quad \rightarrow \quad S_1^A(E_{N,1}) \quad \rightarrow S_2^A(E_{N,1}) \rightarrow\\
&\qquad\qquad\qquad\quad \vdots\\
&S_2^B(E_{N,i}) \rightarrow S_1^B(E_{N,i}) \rightarrow \quad E_{N,i}\\
&\qquad\qquad\qquad\quad \rightarrow \quad S_1^A(E_{N,i}) \quad \rightarrow S_2^A(E_{N,i}) \rightarrow\\
&\qquad\qquad\quad \cdots \quad \rightarrow \quad \cdots \qquad \rightarrow\\
&\qquad\qquad\quad \cdots \quad \rightarrow \quad E_N
\end{aligned}
$$

When conversation participant B is speaking in the intermittent conversation mode, changes in participant B's utterance goals interfere with listener participant A's ability to consistently execute utterance understanding of participant B's utterance content in $S_2^B(E_{N+1})$. This is because the goal structure activated by conversation participant B in $S_2^B(E_N)$—which was active at the start of the utterance—smoothly connects to the Multi-Dimensional Memory Frame activated by conversation participant A in $S_2^B(E_{N-1})$ while A is processing the utterance. However, as the utterance progresses, this structure gets updated, forcing conversation participant A to follow accordingly.

B. Synchronization in Conversation

1) Synchronization Among Conversation Participants in Conversational Behavior: The utterance content of conversation participant B in event E_N reflects the content of participant B's Multi-Dimensional Memory Frame activated in $S_2^B(E_N)$. This includes an evaluation of the results of participant A's utterance in event E_{N-1} and participant B's own utterance in event E_{N-2}. The content of Speaker A's utterance reflects the content of Speaker A's Multi-Dimensional Memory Frame activated in $S_2^B(E_{N-1})$.

In this way, the Multi-Dimensional Memory Frame possessed by each speaker shifts as the conversation progresses, with the activated domain changing based on the evaluation of both the content of the other's utterances and the results of their own utterances.

Verbal behaviors occur by utilizing the hierarchically structured knowledge network shown in Figure 4. Given this, the necessary condition for a smooth conversation to proceed, where speakers alternately process each other's utterances via their respective System-1-Before-Event-Mode without intervention from System 2, can be summarized as follows.

> The activation pattern of Multi-Dimensional Memory Frame via $S_2^B(E_{i-1}) \approx$ The activation pattern of Multi-Dimensional Memory Frame via $S_2^B(E_i)$

This indicates that the activation pattern of one's own Multi-Dimensional Memory Frame that triggers event E_i related to one's own utterance significantly overlaps with the activation pattern of the other's Multi-Dimensional Memory Frame that triggers event E_{i-1} related to the other's utterance. When this condition is met, we can say that conversational behaviors are proceeding synchronously among the conversation participants.

2) Synchronization with VR Systems: The synchronization of conversations involving turn-taking between humans differs in nature from the synchronization that occurs when users interact with VR systems. Dinet et al. [10] identified weak synchronization between the user and the system as the condition for users to interact with multimodal systems with a sense of immersion. Weak synchronization is achieved by designing the specific VR content of the interaction at T_N based on cognitive processing in $S_2^B(E_N),\ S_1^B(E_N),\ S_1^A(E_N),\ S_2^A(E_N)$ regarding the interaction event E_N where the user utilizes the system. To do so, it is necessary to appropriately estimate the propagation of activation within the Multi-Dimensional Memory Frame and the updates to the Multi-Dimensional Memory Frame during the period $[T_N - \beta,\ T_N + \alpha]$.

C. Verbal Behavior and GOMS

This section examines how each element of GOMS relates to conversational behavior. It demonstrates that System 1 and System 2 participating in conversational behavior in a balanced and appropriate proportion is crucial for understanding conversational behavior as it occurs in the real world.

1) The Balance Between Goals and Methods: When we focus on actions that are repeated routinely, these actions manifest based on the GOMS connection structure shown in Figure 4. The individual G, O, M, S shown in Figure 4 are nodes within the Multi-Dimensional Memory Frame corresponding to the knowledge elements necessary to select and execute appropriate actions without continuously referencing the Perceptual-Multi-Dimensional Memory Frame, which is activated through P-resonance with environmental information. Furthermore, action selection and execution must occur in synchronization with an environment whose state changes moment by moment. Therefore, it is reasonable to assume an upper bound exists on their total number. The total number of goals is denoted as $\hat{G}$, total number of methods as $\hat{M}$, total number of selection rules as $\hat{S}$, total number of operators as $\hat{O}$, average depth of the hierarchy as $\bar{N}$, and upper bound on the number of nodes as $\hat{C}$, which is a constant value.

A key feature of the GOMS connection structure is that, due to the finite processing capacity of the brain, either System-2-After-Event-Mode or System-1-After-Event-Mode becomes dominant [9][23]. Depending on the degree of dominance, the following four cases can be considered.

- Case 1: When System-1-After-Event-Mode is dominant, $\hat{M} \gg \hat{G}$ holds.

- Case 2: When System-2-After-Event-Mode is dominant, $\hat{G} \gg \hat{M}$ holds.
- Case 3: When actions occur almost exclusively under System-1-After-Event-Mode, $\hat{M} \gg \hat{G} \sim 0$ holds.
- Case 4: When actions occur almost exclusively under System-2-After-Event-Mode, $\hat{G} \ggg \hat{M} \sim 0$ holds.

2) Understanding Conversational Behavior Through Behavioral Linguistics: The balance between System-1-After-Event-Mode- and System-2-After-Event-Mode-dominance changes depending on the range of communities that the individual is directly and indirectly involved in during their life.

Case 1 corresponds to a smooth conversation mode where one speaker dominates the conversation. The prerequisite for establishing this mode is that the participants share a set of conversational methods. In such behavioral ecology, possessing a set of methods specialized for the situations encountered allows for a perfectly smooth life. Therefore, $\hat{M} \gg \hat{G}$ holds. Since methods, which are cognitive elements, are executed unconsciously, actions driven by System 1 become predominant. In Case 2, each speaker's turn is short, requiring frequent speaker changes to maintain mutual understanding and continue the conversation. When a group consists of both direct communities and indirect societies and/or communication occurs through structural language, System-2-After-Event-Mode becomes the dominant behavioral ecology, resulting in $\hat{G} \gg \hat{M}$. In conversation, reasoning using knowledge stored in the Word- and Relation-Multi-Dimensional Memory Frame have an important role. Flexible adaptation to diverse and changing circumstances is made possible by allocating resources to deliberate System-2-Before-Event-Mode utilizing the goal structures.

Case 1 and Case 2 correspond to conversational behavior in the real-world scenario depicted in Figure 3. The above understanding of these cases was made possible by incorporating in detail how the dual-process—comprising System 1 and System 2, central concepts in behavioral economics—relates to conversational behavior. Therefore, the approach to understanding the verbal behavior demonstrated in this study can be termed behavioral linguistics.

Cases 3 and 4 correspond to ways of understanding that differ from the understanding of conversational behavior proposed in this study. Case 3 involves situations where the goal is extremely limited (such as simply maintaining a conversation). This reduces to stimulus-response behavior that can be executed without cognitive processing, such as producing utterances that can respond to the other person's speech. This aligns with the understanding of verbal behavior based on Skinner's behavioral psychology. Case 4 represents a situation where the methods become extremely limited. The goals determine the details of the utterance. The sequence length of operators leading to the method that specifies the sequence of words to be uttered is short. Consequently, utterances become possible through a large number of goals and their combinations. This aligns with the Chomskyan understanding of linguistic behavior, which posits that goals are symbols and that linguistic actions arise through the manipulation of these symbols.

V. Conclusion and Future Work

This paper proposed a behavioral linguistics theory to explain real-time language generation in everyday conversation. This approach was based on the MHP/RT [6][7] and the dual-process theory of cognition [4][8][15]. The analysis modeled fluent conversation as a PCM cycle alternating between the speaker and the listener, describing in detail how the Multi-Dimensional Memory Frame and GOMS are used for understanding, preparing, and producing utterances. The study also differentiated between fluid and intermittent conversation modes and concluded that this approach, integrating behavioral psychology and behavioral economics, offers a richer framework than models based solely on Skinner [1] or Chomsky [11].

From a theoretical point of view, our paper argued that real-time linguistic behavior cannot be fully captured by static grammatical models or by purely associative accounts of verbal behavior. Instead, it requires a dynamic framework that explains how linguistic choices emerge from moment-to-moment cognitive constraints, environmental cues, and interactive demands. By positioning language production within the PCM cycle, the theory emphasized that utterances are not pre-constructed entities retrieved whole from memory but are assembled progressively through rapid iterations of perception, interpretation, planning, and articulation.

Furthermore, the proposed model highlighted the role of prediction and anticipation in conversation. Speakers and listeners continuously forecast each other's intentions, adapt to turn-taking cues, and adjust their linguistic formulations based on cognitive load and situational incentives. This predictive loop is shaped not only by linguistic competence but also by heuristics, biases, and cost–benefit evaluations, central themes in behavioral economics. As a result, language use is portrayed as an activity controlled by bounded rationality, optimized under real-time processing constraints rather than idealized grammatical rules.

Finally, with our paper, we suggested that a behavioral-linguistic theory grounded in cognitive architecture offers a more comprehensive explanation of conversational competence. It bridged gaps between psycholinguistics, cognitive psychology, and behavioral science, providing a unified account of how humans understand and produce language in everyday interaction. When we understand that speech is built moment-by-moment using limited attention and working memory, we become more aware of why misunderstandings happen. This can help people speak more clearly, listen more actively, and manage turn-taking more smoothly in conversations.

From an applied perspective, several implications can be drawn for our daily lives. Research to realize these implications must be conducted in the future:

(a) Improved learning and teaching. A theory that explains how language is processed in real time can improve language teaching. Teachers can design exercises that match how the brain naturally organizes and retrieves language, making learning more intuitive and efficient;

(b) Reduced communication stress. Knowing that hesitations, pauses, or "ums" are natural results of cognitive processing—not signs of incompetence—can help people feel less anxious when speaking. This is especially helpful for public speaking, second-language use, or social anxiety;
(c) Better HCI. If we understand how humans generate language under time pressure, we can design voice assistants, chatbots, and AI systems that interact more naturally. The theory can guide systems to adapt to human pacing, prediction patterns, and conversational rhythms;
(d) More effective teamwork and decision-making. In workplaces, communication failures are often cognitive failures. Understanding how System 1 and System 2 influence what we say can help people catch biases, avoid rushed judgments, and communicate more thoughtfully in meetings or negotiations;
(e) Insights for therapy and mental health. Speech disruptions often reflect cognitive overload, stress, or emotional pressure. A behavioral model of real-time language can help psychologists better understand how anxiety, ADHD, or trauma affect communication—and help people manage these effects;
(f) Conflict prevention and smoother social interactions. Recognizing that people often speak using quick, automatic processing (System 1) can make us more tolerant of minor errors or emotional reactions in others. It encourages patience and gives a more compassionate understanding of how real conversations work.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI (Grant Numbers 19K12246 / 20H04290 / 22K12284 / 23K11334) and the National University Management Reform Promotion Project.

REFERENCES

[1] B. F. Skinner, *Verbal Behavior*. Appleton-Century-Crofts., 1957.
[2] P. N. Chase, D. W. Ellenwood, and G. Madden, "A Behavior Analytic Analogue of Learning to Use Synonyms, Syntax, and Parts of Speech", *The Analysis of Verbal Behavior*, vol. 24, no. 1, pp. 31–54, 2008. DOI: 10.1007/BF03393055[retrieved: February,2026].
[3] H. A. Simon, "Rational choice and the structure of the environment", *Psychological Review*, vol. 63, pp. 129–138, 1956.
[4] D. Kahneman, "A perspective on judgment and choice", *American Psychologist*, vol. 58, no. 9, pp. 697–720, 2003.
[5] M. Kitajima and M. Toyota, "Simulating navigation behaviour based on the architecture model Model Human Processor with Real-Time Constraints (MHP/RT)", *Behaviour & Information Technology*, vol. 31, no. 1, pp. 41–58, 2012. DOI: 10.1080/0144929X.2011.602427[retrieved:February,2026].
[6] M. Kitajima and M. Toyota, "Decision-making and action selection in Two Minds: An analysis based on Model Human Processor with Realtime Constraints (MHP/RT)", *Biologically Inspired Cognitive Architectures*, vol. 5, pp. 82–93, 2013, ISSN: 2212-683X. DOI: http://dx.doi.org/10.1016/j.bica.2013.05.003[retrieved:February,2026].
[7] M. Kitajima, *Memory and Action Selection in Human-Machine Interaction*. Wiley-ISTE, 2016, ISBN: 9781848219274.
[8] J. S. B. T. Evans, "Dual-processing accounts of reasoning, judgment, and social cognition", *Annual Review of Psychology*, vol. 59, no. Volume 59, 2008, pp. 255–278, 2008, ISSN: 1545-2085. DOI: https://doi.org/10.1146/annurev.psych.59.103006.093629[retrieved:February,2026].
[9] M. Kitajima, M. Toyota, J. Dinet, and K. T. Nakahira, "Transforming Conscious Goals into Unconscious Actions in Real-world Interactions: Real-world Use of Behavioral Ecological Memes via GOMS", *International Journal On Advances in Intelligent Systems*, vol. 18, no. 3 & 4, pp. 173–186, 2025.
[10] J. Dinet and M. Kitajima, "Immersive interfaces for engagement and learning: Cognitive implications", in *Proceedings of the 2015 Virtual Reality International Conference*, ser. VRIC '18, Laval, France: ACM, 2018, 18/04:1–18/04:8, ISBN: 978-1-4503-3313-9. DOI: 10.1145/3234253.3234301[retrieved:February, 2026].
[11] N. Chomsky, *Language*, vol. 35, no. 1, pp. 26–58, 1959.
[12] Y. Barnes-Holmes, S. C. Hayes, D. Barnes-Holmes, and B. Roche, "Relational frame theory: A post-skinnerian account of human language and cognition", in *Advances in Child Development and Behavior*, ser. Advances in Child Development and Behavior, H. W. Reese and R. Kail, Eds., vol. 28, JAI, 2002, pp. 101–138. DOI: https://doi.org/10.1016/S0065-2407(02)80063-5[retrieved:February,2026].
[13] H. Simon, "A Behavioral Model of Rational Choice", *The Quarterly Journal of Economics*, vol. 69, no. 1, pp. 99–118, 1955.
[14] H. A. Simon, *Models of man; social and rational.* Oxford, England: Wiley, 1957.
[15] D. Kahneman, *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux, 2011.
[16] K. E. Stanovich and R. F. West, "Individual differences in reasoning: Implications for the rationality debate?", *Behavioral and Brain Sciences*, vol. 23, no. 5, pp. 645–665, 2000. DOI: 10.1017/s0140525x00003435[retrieved:February,2026].
[17] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language", *Frontiers in Psychology*, vol. Volume 6 - 2015, 2015, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2015.00731[retrieved:February,2026].
[18] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation", in 4, vol. 50, Linguistic Society of America, 1974, pp. 696–735.
[19] S. K. Card, T. P. Moran, and A. Newell, *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1983.
[20] M. Kitajima et al., "Basic Senses and Their Implications for Immersive Virtual Reality Design", in *AIVR 2024 : The First International Conference on Artificial Intelligence and Immersive Virtual Reality*, 2024, pp. 31–38.
[21] M. Kitajima, M. Toyota, and K. T. Nakahira, "Addressing the Symbol Grounding Problem in VR", in *AIVR 2025 : The Second International Conference on Artificial Intelligence and Immersive Virtual Reality*, 2025, pp. 56–62.
[22] M. Kitajima, M. Toyota, and K. T. Nakahira, "Why the Symbol Grounding Problem Matters in Virtual Reality: A Meme-Focused Solution Based on the Model Human Processor with Real-Time Constraints", *International Journal On Advances in Intelligent Systems*, vol. 18, no. 3 & 4, pp. 162–172, 2025.
[23] M. Kitajima, M. Toyota, J. Dinet, and K. T. Nakahira, "Implementation of Structured Memes into Behavioral Ecology via GOMS", in *COGNITIVE 2025 : The Seventeenth International Conference on Advanced Cognitive Technologies and Applications*, 2025, pp. 6–16.