Does the spatial distribution of a speaker's gaze and gesture impact on a listener's comprehension of discourse?

Kazuki Sekine[1], Tomoha Kajikawa[1]

[1] Waseda University
ksekine@waseda.jp

Abstract

This study investigated the impact of a speaker's gaze direction on a listener's comprehension of discourse. Previous research suggests that hand gestures play a role in referent allocation, enabling listeners to better understand the discourse. The current study aims to determine whether the speaker's gaze direction has a similar effect on reference resolution as co-speech gestures. Thirty native Japanese speakers participated in the study and were assigned to one of three conditions: congruent, incongruent, or speech-only. Participants watched 36 videos of an actor narrating a story consisting of three sentences with two protagonists. The speaker consistently used hand gestures to allocate one protagonist to the lower right and the other to the lower left space, while directing her gaze to either space of the target person (congruent), the other person (incongruent), or no particular space (speech-only). Participants were required to verbally answer a question about the target protagonist involved in an accidental event as quickly as possible. Results indicate that participants in the congruent condition exhibited faster reaction times than those in the incongruent condition, although the difference was not significant. These findings suggest that the speaker's gaze direction is not enough to facilitate a listener's comprehension of discourse.

Index Terms: co-speech gestures, eye gaze, discourse

1. Introduction

Human communication is a complex process that involves multiple modalities, including not only speech but also nonverbal cues such as hand gestures and eye gaze. This is particularly true for discourse, which is a linguistic structure that extends beyond a single sentence. A prime example of this is the transmission of events that involve multiple people to a third person. The speaker begins by sequentially introducing the main protagonists sequentially through speech while also locating them in space with gestures and visual references. For example, the speaker may use their right hand to represent protagonist A and their left hand to represent protagonist B. Once the protagonists are assigned to the right and left spaces, the speaker will use these spaces each time they mention the protagonists by pointing to their respective location. Through the combined use of acoustic and visual modalities, the speaker creates cohesiveness in the discourse. This multimodal discourse construction has been previously observed in studies (Gullberg, 2006; McNeill, 2005). The present study aimed to investigate how the use of space, through gestures and eye gaze, during such multimodal discourse construction impacts the discourse comprehension of listeners.

This study focused on Japanese discourse. Unlike English discourse, Japanese discourse is characterized by the frequent omission of the subject. In English, pronouns are obligatory when referring back to a previously introduced subject, but in Japanese, the use of pronouns is not mandatory, and the subject noun phrase is either repeated with a proper noun or omitted when it is re-referred to. The omitted part of the subject is called the zero pronoun, and the part referred to or alluded to in identifying the referent of the zero pronoun is known as the antecedent. The process of identifying the antecedent from the sentence or context in which the zero pronoun is used is called reference resolution. Reference resolution can be categorised into two types: inter-sentence and intra-sentence, depending on whether the zero pronoun and the antecedent are in the same sentence or different sentences. Consider the following two sentences as examples of inter- and intra-sentence reference resolutions, respectively: "Because Peter had a fever, (X) missed his sports activities" and "Because Peter had a fever. (X) missed his sports activities." In both examples, the subject "Peter" is the antecedent, and (X) indicates an omitted subject as a zero pronoun. In English discourse, the zero pronoun would be replaced by the pronoun "he" to refer to "Peter". It is worth noting that in the case of inter-sentence reference resolution, the sentence containing the antecedent and the sentence containing the zero pronoun are not necessarily consecutive.

Research on referential expressions in Japanese discourse has revealed that noun phrases are frequently used to refer to protagonists when they are first introduced or when referring to someone who differs from the subject in the immediately preceding sentence. Conversely, subject omission is commonly employed when referring to the same person as the subject in the immediately preceding sentence (Clancy, 1992; Sekine & Furuyama, 2010).

The use of pronouns and omission of subjects within discourse contributes to coherence. In a discourse where old and new information are intertwined, these linguistic cues may signal the continuation of the topic. Previous studies have demonstrated that coherence in discourse is facilitated not only by speech but also by co-speech gestures (McNeill, 2005; So, Kita, & Goldin-Meadow, 2009). In terms of speakers, Gullberg (2006) found that they consistently use hand gestures and spaces linked to specific objects as a way of creating coherence in discourse. Regarding listeners, research has shown that comprehension of discourse is achieved through the integration of information from two modalities: the speaker's speech and gestures (Goodrich Smith & Hudson Kam, 2012; Hudson Kam & Goodrich Smith, 2011). For example, Sekine and Kita (2015) compared the development of multimodal comprehension of discourse by Japanese children aged 5, 6, and 10 years with that of adults. In Sekine and Kita's (2015) study, a narrative was constructed with three short sentences accompanied by gestures, and a video featuring a female actor creating a story consisting of three short sentences with gestures was employed as a stimulus. The first sentence introduced two protagonists with proper nouns, while the second sentence depicted the actions of the two protagonists using proper nouns, and the third sentence used subject omission to describe the event. The first and second sentences consistently assigned each protagonist to a particular space, one on the left and the other on the right, while the third sentence used a gesture made in either the left or right space, which is referred to as a "reference gesture" (e.g., Sekine & Furuyama, 2010). By examining the gesture used in the third sentence, participants were able to identify the one of the protagonists involved in an accidental event. In the experiment, participants were asked to identify the protagonist involved in the event described in the third sentence. The result showed that participants aged 6 and above were able to do so correctly beyond chance. Sekine and Kita (2017) conducted a study to investigate whether information from gestures can affect the comprehension of a subsequent sentence even after the gestures have disappeared. Specifically, they examined whether information conveyed by reference gestures in the first and second sentences affects the comprehension of a subsequent third sentence that was produced without gestures. The stories and structures of the stimulus sentences were similar to those used in their previous study (Sekine & Kita, 2015), but with the second sentence describing the actions of each person and the third sentence indicating the person to whom the event occurred, even if only through spoken information. The experimental participants were asked to identify the person involved in the third sentence, and the results showed that listeners maintained the reference gesture in the first and second sentences when comprehending the subsequent third sentence, even after the gesture had ended. Regarding the facilitating effect of gestures on discourse comprehension, Gunter and Weinbrenner (2017) investigated the role of gestures in discourse comprehension using EEG. The results showed that when the frequency of trials in which the gesture was a cue for the reference resolution and the frequency of trials in which it was not a cue were both 50%, the participants did not use the gesture as a cue for discourse comprehension. However, when the gesture was consistently a useful cue, participants did utilise it for discourse comprehension. Therefore, studies examining the role of gestures in discourse comprehension demonstrated that the speaker's gestures during discourse have a significant impact on creating coherence and influencing the reference resolution performed by the listener.

Several studies suggest that eye gaze can contribute to discourse coherence and enhance the listener's comprehension of the discourse. In communication by sign language, gaze has been identified as a factor in constructing discourse coherence. Thompson, Emmorey, Kluender, and Langdon (2013) analysed gaze movements in American Sign Language (ASL) conversation and found that gaze is directed toward the sign space, which is the frontal space used for sign production. Specifically, gaze is not used to indicate referents for a person in ASL, but it is used to indicate referents for places related to the person. In audible communication, Laeng, Bloem, D'Ascenzo, and Tommasi (2014) found that gaze movements during perception and recall significantly overlap, and that closer gaze movements during recall than during perception are associated with better performance on a spatial memory task. Therefore, it can be concluded that gaze plays a role in processing spatial information.

Hence, while the role of gaze in constructing discourse has been explored in sign language, and gaze patterns during perception and recall have been investigated for those who can hear, its use in conjunction with space in discourse has not been thoroughly studied for those who can hear. In studies that examine the connection between discourse and gesture through stimulus videos, the speaker in these videos consistently maintains eye contact with the camera and the listener (Sekine & Kita, 2015, 2017). However, in natural conversational settings, the speaker's gaze is not always fixed on the listener.

The aim of the present study is to manipulate the speaker's gaze distribution in space to investigate its influence on the listener's discourse comprehension, particularly its impact on reference resolution. It is important to note that in this study, gaze movements involve head movements given the spontaneity. To achieve this, a video was created featuring a female actor narrating a story composed of three short sentences involving two protagonists, Taro (male) and Kanako (female). In the first two sentences, the speaker consistently assigned each protagonist to either the right or left space through gestures. In the third sentence, the speaker's gaze was directed to either the left or right space. Participants in the experiment were required to watch the video and answer a question about the target protagonist of the third sentence, which contained a zero pronoun. The independent variable was the gaze direction, which consists of three conditions: the congruent condition, where the gaze was directed towards the reference space of the target protagonist; the incongruent condition, where the gaze was directed towards the reference space of the other protagonist who was not the target protagonist; and the speech-only condition, where the speaker stared at the video camera without moving her gaze. The two dependent variables were the proportion of trials with correct responses and the reaction time. The processing performed by the participants in this study was inter-sentence reference resolution.

This study has the potential to provide valuable insights into the field of communication research regarding the multimodal realisation of discourse comprehension. Specifically, it demonstrates how the listener's comprehension of discourse results from the integration of various linguistic and nonverbal cues, including speech, gestures, space, and gaze. Moreover, this research sheds light on the advantages of adopting the speaker's perspective for listeners when participating in conversational exchanges.

The first hypothesis (Hypothesis 1) posits that there exists no noteworthy discrepancy in the proportion of correct responses (accuracy rate) among the congruent, incongruent, and speech-only conditions, because the target protagonist can also be deduced by speech information alone in the third sentence.

The second hypothesis (Hypothesis 2) posits that there will be a facilitating effect of gaze on discourse comprehension, resulting in significantly shorter reaction time in the congruency condition compared to the speech-only condition. According to Gunter and Weinbrenner (2017), the use of gestures by experimental participants, if deemed useful for discourse comprehension, can have a facilitating effect on discourse comprehension. In this study, the congruent, incongruent, and speech-only conditions are implemented using a between-subjects design, assuming that experimental participants view gaze as a helpful cue for the task in the congruent condition, thus shortening their reaction time. Conversely, it is also hypothesized that the incongruent condition will have a disruptive effect, leading to significantly longer reaction time than in the speech-only condition. This is supported by Sekine and Kita's (2017) findings, which suggest that the presentation of information inconsistent with the reference space by gesture can interfere with discourse comprehension.

2. Methods

2.1. Participants

Participants in the study consisted of 30 undergraduate and postgraduate students, including 14 females and 15 males, with one non-response. All participants were native Japanese speakers with a mean age of 21.03 years (SD = 1.56), ranging from 18 to 25 years. Participants reported no issues with audiovisual acuity.

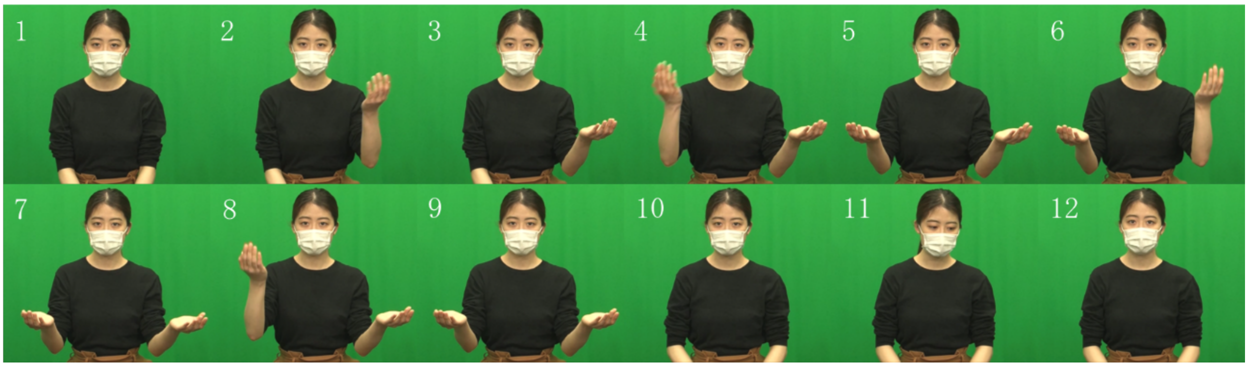
2.2. Experiment design

The experiment employed a two-way mixed design. The first independent factor, referred to as gaze direction, was a between-subjects factor that examined the congruency between the concurrent verbal reference to the target person and the speaker's gaze direction towards the reference space. This factor had three conditions: the congruent condition, where the gaze was directed towards the target person's reference space (Figure 1); the incongruent condition, where the gaze was directed towards another person's reference space; and the speech-only condition, where the gaze was fixed on the camera without any gesture. It is important to note that no gestures were made at all in the speech-only condition. As the reference spaces are created by both gestures and eye gaze in the congruent condition, to see the effect of visual modalities on discourse comprehension, both modalities should be excluded from the speech-only condition as a control group. The second independent factor was block for the stimulus presentation, which is a within-subjects factor. We presented stimuli in three blocks. We included blocks as an independent factor because it was important to examine whether participants noticed that they could identify the target protagonist only by listening to speech as trials went on. Two dependent variables were measured: the proportion of trials with correct responses and the reaction time for correct response trials. Reaction time was defined as the duration between the start of the utterance of the word after 'unfortunately' in the third sentence and the beginning of the participant's verbal response.

2.3. Materials

A total of 39 stories were created for this study, including 23 stories used in Sekine and Kita's (2017) study and additional stories created for this current study. Three out of the 39 stories were used for the practice trial, and the remaining 36 were used in the main trial.

For each story, an audiovisual videoclip was created featuring a female actor narrating the story. In the congruent and incongruent conditions, the actor produced gestures and directed their gaze towards a specific reference space while recounting the stories. Each story was comprised of three short sentences, which has two protagonists, Taro (male) and Kanako (female) (see Figure 1). The male protagonist was always introduced first, followed by the female. The first sentence introduced both protagonists by their names, while the second sentence described their actions along with the corresponding proper nouns. The third sentence mentioned an accidental event that involved one of the protagonists but omitted the subject. The second part of the third sentence contained a cue word, associated with the accidental event (hereafter referred to as the cue word), which allowed the identification of the target protagonist of the third sentence**.** A cue word was always an object name which was related to the accidental event which the target protagonist was involved in. As will be discussed later, during the congruent and incongruent conditions, the actor's gaze would shift towards the reference space of either protagonist before the cue word. In the first sentence, the actor allocated reference space for two protagonists through gesture while introducing the two protagonists. In the second sentence, the actor utilised a gesture within the person reference space.



First sentence: Taro (2,3) and Kanako (4,5) were preparing to go out.

Second sentence: Taro (6,7) was brushing his teeth and Kanako (8,9) was drying her hair.

Third sentence: Unfortunately (11), **the toothpaste** spilled on him.

Figure 1: *An example of a visual stimulus and a short sentence used in the experiment. The numbers in parentheses in the short sentence correspond to the numbers in the pictures where gestures occurred (from 2 to 8) or the gaze shifted (11). This is an example of the congruent condition. The word(s) in bold are the cue word(s) in the story.*

With regard to the allocation of the two protagonist reference spaces by gesture, during the utterance of the proper names in sentences 1 and 2, one protagonist was consistently assigned to the space to the lower right and the other to the lower left. The allocation gestures, depicted in Fig. 1, were produced with an upward and downward motion of the hand, with the palm facing upwards. During the recording of the videoclips, care was taken to ensure that when one hand moved up, the other hand did not move up simultaneously. The gestures were maintained from the time each person was allocated to a space in the first sentence until the end of the second sentence.

Her gaze was redirected from the beginning of the word 'unfortunately' in the third sentence and remained fixed on the person reference space until the end of the cue word. In the congruent condition, the actor shifted her gaze to the target protagonist's reference space during the depiction of the accidental event in the third sentence, out of the two protagonist reference spaces created by gesture in the first and second sentences (as illustrated in Figure 1, number 11). In the incongruent condition, her gaze was directed towards the non-target person reference space which was not the target person. By shifting her gaze before the cue for identifying the target protagonist, participants could potentially resolve reference resolution from the gaze information. In the speech-only condition, the actor's gaze was consistently directed towards the video camera positioned in front of her.

The actor wore a mask to prevent participants from perceiving information from mouth movements, and to add speech sounds, that were separately recorded, to videoclips. Furthermore, although proper nouns were used in the second sentence and the subject was omitted in the third sentence, the omission of subjects is common in Japanese discourse (e.g., Clancy, 1992; Sekine & Furuyama, 2010). Therefore, the discourse structure used in this study was perceived as natural by the participants.

2.4. Experimental equipment

To present the instructional slides and stimulus videos to the participants, we used a high-resolution monitor (24UD58-B, 23.8 inches, LG). We also used a voice response input device (SV-1 Voice Key, Cedrus Corporation) to measure the voice onset for reaction time analysis. The participants wore a headset (an accessory of SV-1), and the experimental stimuli were presented through a software (SuperLab6, Cedrus Corporation). A video camera (Panasonic) was used for monitoring purposes.

2.5. Procedures

The experiments were conducted individually in a soundproof laboratory. Each participant was positioned in a chair facing the monitor, with ad distance of approximately 75cm between them. Participants were assigned one of three conditions: the congruent, incongruent, or speech-only condition.

The experimental task required participants to identify which protagonist, Taro or Kanako, was involved in the accidental event based on the third sentence, and to speak their answer into the microphone for each trial. To guide their responses, participants were instructed to respond quickly and accurately as much as they can, and to speak in a louder tone with particular emphasis on the first syllable. This instruction was necessary to facilitate accurate speech response input.

Participants first received six practice trials with a fixed order, followed 36 experimental trials divided into three blocks. The presentation order of the 36 stories was randomized for each participant. We also counterbalanced the gender and he location (left or right) of the target protagonist.

Each trial began with a fixation cross presented at the center of the screen for one second, followed by the automatic playback of the stimulus video. Once the participant's vocal response was registered, the fixation cross reappeared, and the stimulus video played again. The total duration of the 36 trials was approximately 20 minutes, with each video clip lasting approximately 30 seconds.

For the correct response analysis, we calculated the proportion of correct responses by dividing the number of trials with correct responses by 36 (total trials). For the reaction time analysis, we excluded trials that exceeded the individual mean of ±2 standard deviations for each participant in the correct trials. We then calculated the average reaction time for correct trials and the trials in which participants successfully completed the speech response input. Note that the number of trials used to calculate the correct response rate and reaction time may have differed per participant.
The experiment was approved by the Ethics Review Committee for Research Involving Human Participants at Waseda University (2022-126).

3. Results

3.1. Proportion of trials with correct answers

To investigate whether the speaker's eye gaze influences the listener's correct response, we conducted a two-way mixed analysis of variance (ANOVA) with a 3 (gaze direction: congruent, incongruent, speech-only; between-subjects factor) × 3 (block: block 1, 2, 3; within-subjects factor) experimental design. The dependent variable was the proportion of trials with correct responses. The results showed no significant interaction between gaze direction and block (*F* (4, 54) = .90, *p* = .47), no significant main effect of block (*F* (2, 58) = 0.34, *p* = .71), and no significant main effect of gaze direction (*F* (2, 27) = 1.16, *p* = .33) (Table 1).

Table 1: *Mean and standard deviation for the proportion of trials with correct responses for each condition.*

Condition	Mean	SD
Congruent	0.97	0.03
Incongruent	0.96	0.05
Speech-only	0.98	0.02

3.2. Reaction time

To investigate whether the speaker's eye gaze influences listener's reaction time, we conducted a two-way mixed analysis of variance (ANOVA) with a 3 (gaze direction: congruent, incongruent, speech-only; between-subjects factor) × 3 (block: block 1, 2, 3; within-subjects factor) experimental design. The dependent variable was reaction time. The results showed no interaction between gaze direction and block (*F* (4, 54) = .731, *p* = .57), but a main effect of block (*F* (2, 58) = 19.93, *p* < .001, η2 = .41). Further examination using the LSD method for multiple comparisons showed that reaction time in blocks 1 (1474.86 ms) was significantly shorter than in block 2 (1394.99 ms) (p = .001) and block 3 (1315.92 ms) (p < .001), and reaction time in block 2 was significantly shorter than in block 3 (p = .004). These findings indicate that the reaction time decreased across all gaze direction conditions as the number of trials increased.

We also found a marginally significant result towards a main effect of gaze direction (F (2, 27) = 2.75, p = .08, η2 = .17) indicating that the speaker's gaze direction is not enough to influence the listener's information processing in discourse comprehension. The multiple comparisons (LSD method) revealed a significant difference between the congruent and incongruent conditions (p = .03). The congruent condition (1277.15 ms) showed a shorter reaction time compared to the incongruent condition (1507.36 ms) (Figure 2).

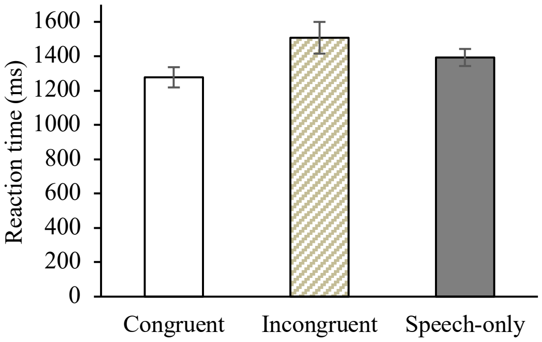


Figure 2: *Reaction time in gaze direction conditions. Error bars indicate standard errors.*

4. Discussion

The current study investigated the impact of the speaker's gaze distribution in reference space on the listener's comprehension of discourse. There are two main findings.

Firstly, the result showed no significant effect of gaze direction on the proportion of correct responses. This outcome supports hypothesis 1, which postulates that there is no significant difference in the proportion of correct responses among conditions. The proportion of trials with correct responses exhibited a ceiling effect. This effect can be ascribed to the characteristic of this task. In the task of the current study, the participants in any conditions could identify the target protagonist from the speech information only, specifically the cue word in the third sentence. Due to this characteristic, all participants equally performed well.

Secondly, the reaction time analysis revealed a significant trend in the main effect of gaze direction. The congruent condition resulted in shorter reaction times compared to the incongruent condition, but no significant differences were found between the congruent and speech-only conditions, nor between the incongruent and speech-only conditions. Thus, this finding contradicts hypothesis 2, which postulates that both the facilitating and interference effects of eye gaze are observed.

Overall, the results of the current study suggest that listeners' comprehension of discourse was not quicker when the speaker's gaze corresponded with her utterance, as opposed to when her gaze was inconsistent with her utterance. This trend partly aligns with that reported in Sekine and Kita (2017) where it was found that gestures hindered discourse comprehension rather than facilitated it. A possible explanation for the lack of the facilitating effect of eye gaze in the present study might be that few participants perceived eye gaze as a crucial cue for the task even if they were aware of the speaker's eye movements. Also, a possible reason for the lack of the interference effect of eye gaze is that many participants ceased to attend to the speaker's nonverbal cues, including eye movements and gestures toward the end of the experiment. Indeed, some participants reported in the post-experiment questionnaire that they noticed that they could attain the correct answer only by listening to speech. Thus, as a future task, it is deemed necessary to design a task that compels participants to fixate on the display monitor on which the video stimuli are displayed. It would facilitate participant's attentive engagement with the speaker's gestures and eye gaze.

Laeng et al. (2014) reported that higher performance on a spatial memory task was achieved when gaze movements during image recall were more similar to those during stimulus perception. It is conceivable that bringing the listener's gaze to the cued person reference space by means of the speaker's gaze distribution may have prompted the listener to recall the name of the target person assigned to that person reference space. Nevertheless, as the present study did not reveal the presence or absence of a facilitating or interfering effect of gaze on discourse comprehension, it is equally possible that directing the listener's gaze to the wrong person reference space may have interfered with the listener's recall of the name of the person who was the correct answer. Thus, investigating the relationship between the listener's memory performance and the speaker's gaze direction would be another important future task.

5. Conclusions

We have concluded that the speaker's eye gaze, directed towards the reference space that was established by gestures, does not strongly influence the listener's comprehension of discourse. This finding provides valuable insight into the multimodal nature of discourse comprehension.

6. References

Clancy, P. M. (1992). Referential strategies in the narratives of Japanese children. *Discourse Processes*, *15*, 441-467.

Goodrich Smith, W. & Hudson Kam, C. L. (2012). Knowing 'who she is' based on 'where she is': The effect of co-speech gesture on pronoun. *Language and Cognition*, *4*, 75-98.

Gullberg, M. (2006). Handling Discourse: Gestures, Reference rracking, and communication strategies in early L2. *Language Learning*, *56*, 155-196.

Gunter, T. C., & Weinbrenner, J. E. D. (2017). When to take a gesture seriously: On how we use and prioritize communicative cues. *Journal of Cognitive Neuroscience*, *29*, 1355-1367.

Hudson Kam, C. L., & Goodrich Smith, W. (2011). The problem of conventionality in the development of creole morphological systems. *The Canadian Journal of Linguistics*, *56,* 109-124.

Laeng, B., Bloem, I. M., D'Ascenzo, S., & Tommasi, L. (2014). Scrutinizing visual images: the role of gaze in mental imagery and memory. *Cognition*, *131*, 263-283.

McNeill, D. (2005). *Gesture and thought.* Chicago: University of Chicago Press.

Sekine, K., & Furuyama, N. (2010). Developmental change of discourse cohesion in speech and gestures among Japanese elementary school children. *Revista di Psicolinguistica Applicata*, *10*, 97-116.

Sekine, K., & Kita, S. (2015). Development of multimodal discourse comprehension: cohesive use of space by gestures. *Language, Cognition and Neuroscience*, *30*, 1245-1258.

Sekine, K., & Kita, S. (2017). The listener automatically uses spatial story representations from the speaker's cohesive gestures when processing subsequent sentences without *Acta Psychologica*, *179*, 89-95.

So, W, C., Kita, S., & Goldin-Meadow, S. (2009). Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive Science*, *33*, 115-125.

Thompson, R. L., Emmorey, K., Kluender, R., & Langdon, C. (2013). The eyes don't point: Understanding language universals through person marking in American Signed Language. *Lingua*, *137*, 219-229.