Acta Psychologica 179 (2017) 89–95

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

The listener automatically uses spatial story representations from the speaker's cohesive gestures when processing subsequent sentences without gestures

Kazuki Sekine[a,b,*], Sotaro Kita[c]

[a] Radboud University, The Netherlands
[b] Max Planck Institute for Psycholinguistics, The Netherlands
[c] University of Warwick, United Kingdom

ARTICLE INFO

Keywords:
Gesture
Speech
Discourse
Multimodal communication
Speech comprehension
Speech-gesture integration

ABSTRACT

This study examined spatial story representations created by speaker's cohesive gestures. Participants were presented with three-sentence discourse with two protagonists. In the first and second sentences, gestures consistently located the two protagonists in the gesture space: one to the right and the other to the left. The third sentence (without gestures) referred to one of the protagonists, and the participants responded with one of the two keys to indicate the relevant protagonist. The response keys were either spatially congruent or incongruent with the gesturally established locations for the two participants. Though the cohesive gestures did not provide any clue for the correct response, they influenced performance: the reaction time in the congruent condition was faster than that in the incongruent condition. Thus, cohesive gestures automatically establish spatial story representations and the spatial story representations remain activated in a subsequent sentence without any gesture.

People often produce gestures while speaking. Research on such co-speech gestures has revealed that the listener/observer can take up information from the speaker's gestures and use it to comprehend an underlying overall message (Beattie & Shovelton, 1999; Cassell, McNeill, & McCullough, 1999; Kendon, 1994). This study examined whether a listener uses spatial information expressed in gestures even after the gesture has disappeared.

Most of the previous research on gesture comprehension focused on the processing of a single gesture at word or sentence level. For example, some studies shown that adults and children can pick up information conveyed exclusively in gestures (e.g., Broaders & Goldin-Meadow, 2010; Kelly & Church, 1998; Namy, Campbell, & Tomasello, 2004; Morford & Goldin-Meadow, 1992). Other studies reported that adults and children integrate gesture and speech, each of which contributes unique information to the unified interpretation (e.g., adults: Cocks, Sautin, Kita, Morgan, & Zlotowitz, 2009; Kelly, Özyürek, & Maris, 2010; children: Kelly, 2001; Sekine, Sowden, & Kita, 2015). Thus, the findings from these studies suggest that adults and children can pick up information conveyed by gesture and integrate it with information from the concurrent speech. However, these studies have focused on comprehension of speech and a single gesture, and thus integration of speech and gesture at word or sentence level. Comprehension of speech and a sequence of gestures at the discourse level is under-studied.

Studies on gestures in discourse have revealed that during a narrative, an adult speaker builds coherent discourse by using linguistic devices and speech-accompanying gestures (Gullberg, 2006; McNeill, 2005; McNeill & Levy, 1993; Yoshioka, 2005). As McNeill (1992) argued, adult speakers often use gestures to indicate continuity of a topic by repeating the same form or the same location. For example, when a new protagonist is introduced in a story, an adult speaker locates the referent to specific space in front of him or her by a pointing gesture or an iconic gesture (iconic gestures are gestures that depict objects, actions and movements on the basis of similarity). When mentioning the same referent again later, (s)he gesturally indicates the same location (Gullberg, 2006; So, Kita, & Goldin-Meadow, 2009). McNeill and Levy (1993) argued that the assigned spaces for referents were gestured more frequently when characters were re-introduced with explicit referring form such as a noun phrase than when the narrative maintained focus on one character with a less explicit form such as a pronoun. So et al. (2009) found that speakers tended to produce gestures in a particular location in gesture space to identify referents that were also uniquely specified in speech (e.g., two different gender protagonists were referred to by the pronouns "he" and "she"), rather than referents that

[*] Corresponding author at: Max Planck Institute for Psycholinguistics, Room 308, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands.
E-mail address: kazuki.sekine@mpi.nl (K. Sekine).

http://dx.doi.org/10.1016/j.actpsy.2017.07.009
Received 30 September 2016; Received in revised form 10 July 2017; Accepted 17 July 2017
Available online 24 July 2017
0001-6918/ © 2017 Elsevier B.V. All rights reserved.

were ambiguous in speech (e.g., two same gender protagonists were referred to by the same pronoun "he"). These findings suggest that speakers tend to use locations in gesture space to indicate referents that are lexically specified by the concurrent speech.

Gestural reference-tracking is attained when specific gestural behaviors, whose features (e.g., location, handedness, movement, orientation, hand shape) are repeated, are systematically associated with referential expressions in speech (Gullberg, 2006). Such an association establishes explicit, visual co-reference, and thus, enhances the cohesiveness of the discourse (McNeill, 2005). Previous studies have shown that referential space is constructed and used not only by individuals (So et al., 2009) but also by conversation participants (Stec & Huiskes, 2014). Gestures used in establishing locations for referents and tracking the references in the discourse are called cohesive gestures (McNeill, 1992), as they contribute to the discourse cohesion.

There have only been seven studies that investigated how listeners process cohesive gestures. Two studies investigated how spatial information in cohesive gesture influenced subsequent speech and gesture production. First, Cassell et al. (1999) showed that listeners take up information from the cohesive use of space in gesture. Cassell et al. presented a video-recorded narrative to adult participants, who then retold the story to a listener. In the stimulus narrative, a narrator located two protagonists in his frontal space with deictic gestures, and then linguistically referred back to one of the protagonists while pointing to the wrong space (the space for the other referent). When retelling the narrative, participants incorporated information from gesture and speech even when they were incongruent with each other.

Second, an EEG study by Gunter, Weinbrenner, and Berndt (2012) found that the brain prepares to produce cohesive gestures even when producing the gestures was not required for the task. Participants watched video clips where a narrator tells stories and establishes two locations for two referents (e.g., left side for cats and right side for dogs) with cohesive gestures. Then the participants were asked to respond verbally to a question like "Which animal barks?" (no gesture was produced). Event related potentials (ERPs) revealed that they covertly prepared to produce a cohesive gesture by the left or right hand that was compatible with the location (left or right) where the gestures in the stimulus placed the relevant referent.

Five previous studies investigated how cohesive gestures influence comprehension of discourse, which is the topic of the current study. Three EEG studies (Gunter & Weinbrenner, 2017; Gunter, Weinbrenner, & Holle, 2015; Weinbrenner, 2017) showed that cohesive gestures influenced listener's comprehension of a sentence even when the gesture was not crucial for interpreting the sentence. In Gunter et al.'s (2015) study, participants were presented with video clips of an interview between an interviewer and an interviewee. In each video clip, the interviewee talked about a topic consisting of two opposing referents (e.g., "Donald vs. Mickey") and assigned the two referents to two locations in gesture space with cohesive gestures (e.g., left space for Donald and right space for Mickey). With the target sentence at the end of each topic, the interviewee produced a cohesive gesture that was either congruent (pointing to the left while saying "Donald") or incongruent (pointing right while saying "Donald") to the previously established location with a sentence like "As far as I know, Donald was created later". The target sentence was unambiguous and fully interpretable without the accompanying gesture. Participants were asked to pay attention to the video clips as they were given a memory task about the video content, which was neither about the content of the target sentence nor was it related to gesture. The result showed that the congruency between speech and cohesive gesture influenced ERPs recorded from the participants: N400 and P600 components were larger in the incongruent condition than in the congruent condition. This indicates that it was more difficult to process the interviewee's message when the cohesive gesture was incongruent with the speech, even though participants were not asked to pay attention to gestures while watching the stimulus videos. The same pattern of results was found by other studies (Gunter & Weinbrenner, 2017; Weinbrenner, 2017).

Fourth, Goodrich Smith and Hudson Kam (2012) showed that cohesive gestures can influence the interpretation of otherwise ambiguous sentences. They investigated whether cohesive gestures influenced interpretation of ambiguous pronouns. Participants watched video clips of a narrator telling stories that ended with a sentence with an ambiguous pronoun: e.g., "*Annie and Sarah are having a picnic in the park. They have a lot of food with them. Annie is carrying the picnic basket, and Sarah has a blanket to sit on. She is excited about the cookies*". In the first two sentences, the narrator consistently located the two protagonists to either her right or left side with cohesive gestures. In the last sentence, the narrator's gesture was manipulated: she either produced no gesture, indicated the location of the first-mentioned protagonist or the second-mentioned protagonist. After each clip, the participants were presented with a question (without any gestures) about the referent of the ambiguous pronoun (e.g., "*Who is excited about the cookies?*"). The participants tended to respond with the referent that was consistent with the location indicated by the gesture in the third sentence. This indicates that cohesive gestures influence people's interpretation of the ambiguous pronounce.

Fifth, Sekine and Kita's (2015) study showed that cohesive gestures influence comprehension of discourse in elementary school children. They examined how well Japanese 5-, 6-, 10-year-olds and adults integrated information from spoken discourse and cohesive gestures in comprehension. The participants were presented with three-sentence stories. In the first two sentences, a narrator referred to two protagonists by full nouns and an event involving them (e.g., "Nori-kun and Yuuto-kun are crossing a pedestrian bridge. Nori-kun and Yuuto-kun are ascending stairs." Note that Nori-kun and Yuuto-kun are Japanese boys' names). The narrator produced gestures to consistently locate each of the two protagonists in two distinct locations (e.g., left space for Nori-kun and right space for Yuuto-kun). In the third sentence, she described a protagonist's movement without explicitly mentioning any protagonists, which is grammatically possible in Japanese (e.g., "and suddenly, (one) tumbled down"). In addition, she iconically depicted one of the protagonists' movements within the right or left space. Thus, participants could infer which character did the movement only if they took the gestures into account. Then, they were asked to indicate which protagonist performed the action in the third sentence. The result showed that 6- and 10-year-olds, and adults consistently selected the protagonist consistent with the location indicated by the iconic gesture in the third sentence, but not 5-year-olds, whose choice was at chance.

These five studies on the impact of cohesive gestures on discourse comprehension showed that cohesive gestures influence processing of the concurrent sentence. However, it is not clear if cohesive gestures influence processing of a subsequent sentence *without* any accompanying gestures.

Hudson Kam and Goodrich Smith (2011) showed that spatial information encoded in gestures persists beyond the sentence that the gestures co-occurred with. In their study, participants were presented with video clips where an actor located two protagonists in the left and the right side of the gesture space with cohesive gestures (e.g., Andrea on the right and Bobby on the left, from the participant's perspective). After watching each clip, participants chose one of two pictures that best represented the story they heard. One picture showed one protagonist on the right and the other on the left, and the other picture flipped the left-right positions of the two protagonists. Participants systematically picked the picture with the two protagonists located in the left-right positions compatible with where the gestures located the two protagonists (e.g., Andrea dancing on the right, and Bobby singing on the left, from the participant's perspective). However, in this task, because locations indicated by gestures were the only clue that allowed participants to select the response, the task required participants to pay attention to the gestures and try to remember the locations even after the story. Thus, it is still not clear whether the spatial representation created by gestures is activated *automatically*, that is, in a situation

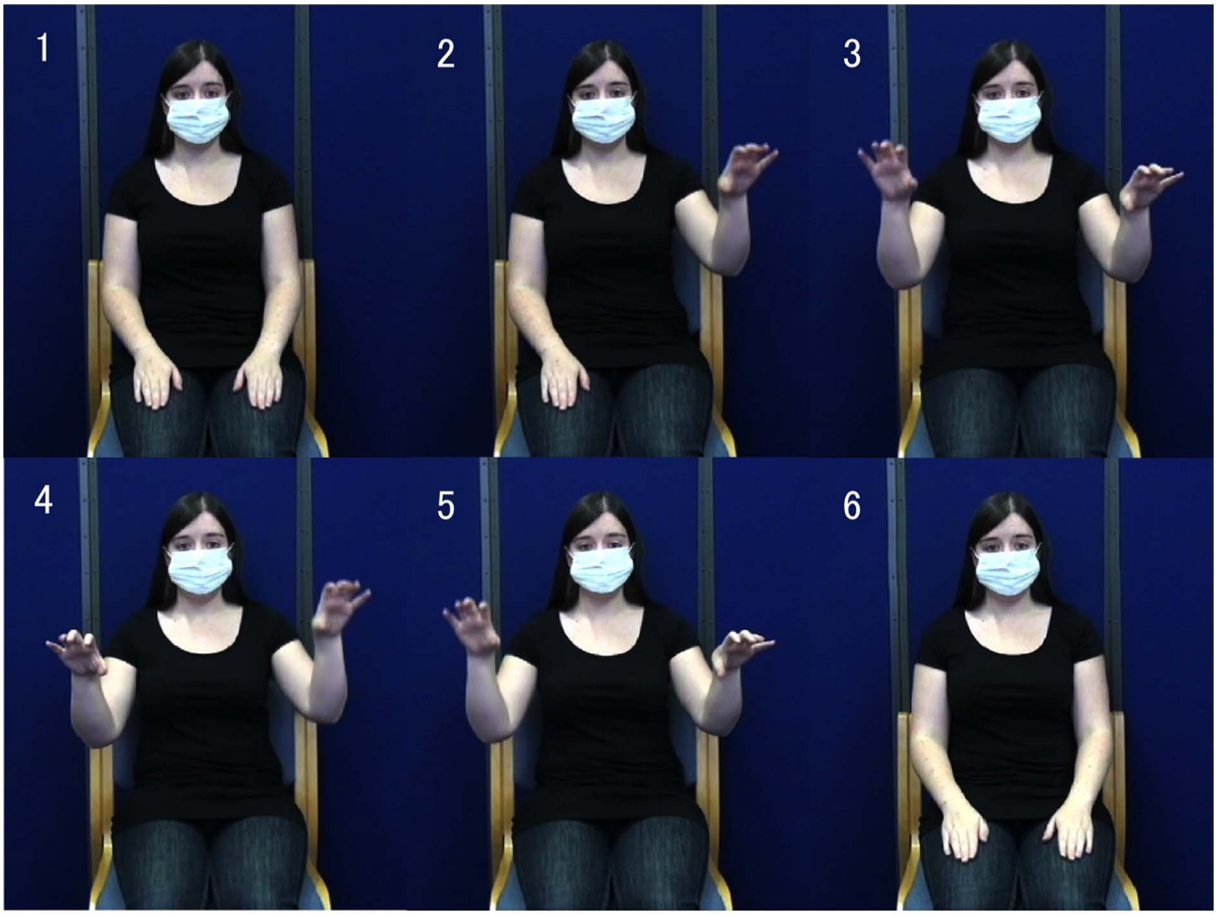

Fig. 1. An example of a visual stimulus and a short sentence used in Experiment 1. The numbers in parentheses in the short sentence correspond to the numbers in the pictures where gestures occurred (from 2 to 5) or the protagonists name appeared (6). This is an example of the incongruent condition (In the congruent condition, the locations of the words "Betty" and "Gary" were flipped.)

(2) **Gary** and (3) **Betty** were preparing to go out.

(4) **He** was brushing his teeth and (5) **she** was drying her hair.

(6) Unfortunately the toothpaste spilled on him.

where the gestures are irrelevant to the task because they do not provide any information about the correct response in the task.

To summarise, the previous literature left the following question open: does the listener automatically activate spatial information of cohesive gestures when interpreting a subsequent sentence without a gesture? The present study investigated this question by conducting three experiments, which manipulate stimulus-response compatibility (Kornblum, Hasbroucq, & Osman, 1990). The stimulus-response compatibility refers to the fact that responses are faster when the stimulus locations (e.g., a stimuli appeared on the right location) are compatible with the locations of their corresponding response keys (e.g., the right-side key) than when they incompatible (e.g., the left side key).

In our three experiments, English-speaking adult participants were presented with video clips where a female actor tells a short story consisting of three sentences. In the first sentence, she introduced two protagonists (one male and one female), and in the second sentence she referred back to them (see Fig. 1). Every time she mentioned the two protagonists, she located the two protagonists in consistent locations (her left and right side) in the gesture space with cohesive gestures. In the third sentence, she described an accident in which one of the protagonists was involved *without* an accompanying gesture. Which protagonist was involved could be inferred from the subject NP and the verb, and became completely unambiguous in the final word. At the beginning of the third sentence, the two protagonists' names were visually presented above the actor's shoulders. The locations of the protagonists' names were either congruent or incongruent with where gestures localized the two protagonists in the preceding discourse. The participants were instructed to indicate which protagonist was in the accident by pressing the key on the computer keyboard that was on the same side as the relevant name on the computer screen. Note that participants could choose the correct response based on the speech information alone, and that the cohesive gestures did not provide any useful information for the judgement. Thus, any influence of gesture on the performance indicates that the gesturally established meanings of spatial locations are automatically activated.

In Experiment 1, every story had the same protagonists, Gary and Betty, and every story located Gary and Betty in the same respective positions (left or right). This allows spatial representations of the two protagonists to build up across trials. The locations of the protagonists' names were either congruent or incongruent with where gestures localized the two protagonists in the preceding discourse. Experiment 2 investigated whether cohesive gestures can leave a strong enough representation after only two sentences (within a trial), in contrast to Experiment 1 in which every story located same protagonists in the same respective positions across trials. We examined this by varying the names of protagonists and the locations of the male/female protagonists for each trial. Experiment 3 examined whether a spatial representation created by cohesive gestures facilitate or interfere with key-press responses by comparing the performance in the congruent and incongruent conditions to the baseline condition in which the entire discourse did not include any gestures.

We tested two competing hypotheses; *Active gestural discourse* *representation hypothesis* and *Semi-active gestural discourse representation hypothesis*. The *Active gestural discourse representation hypothesis* proposes that the spatial story representation created by the speaker's cohesive gestures automatically remains active throughout the discourse, even in sentences without any gestures. According to this hypothesis, the performance in the congruent condition should be better than that in the incongruent condition.

Alternatively, the *Semi-active gestural discourse representation hypothesis* proposes that the spatial story representation created by the speaker's cohesive gestures becomes non-active once cohesive gestures finish and this can be re-activated only with a new gesture. That is, gestural representation becomes only "semi-active" in the sense of Chafe's (1987), and can be activated again when the referent is re-introduced into the discourse by another cohesive gesture. According to this hypothesis, the performances between the congruent and the incongruent condition should not be different.

1. Experiment 1

1.1. Method

1.1.1. Participants

Twenty native English speakers (10 female and 10 male) took part in this study. The participants' age ranged from 18 to 37 years (M = 21.10, SD = 4.39). They reported having normal or corrected-to-normal vision and audition.

1.1.2. Material and apparatus

An actor was filmed producing combinations of gestures and a short passage. All the video stimuli and data can be downloaded from the following URL: https://osf.io/52qjc. Twenty-three stories were made in total (three for practice, twenty for the main experiment). Each story had different events but the protagonists were always Gary and Betty. The lower part of the actor's face was covered by a mask because the speech from separate recordings was edited into the stimuli. The editing was necessary to create two different visually identical versions of video clips that differ only in the last sentence starting from 'unfortunately' (one with Gary, the other for Betty: see an example in Fig. 1).

Each story consisted of three short sentences and gestures. Gestures were accompanied with speech in boldface in the following example. In the first sentence, the actor introduced a male protagonist and a female protagonist with their proper names (e.g., "**Gary** and **Betty** were preparing to go out") in the subject position. In the second sentence, she described two different events that each protagonist was involved in with pronouns (e.g., "**He** was brushing his teeth and **she** was drying her hair"). In the third sentence, which always started with the word "unfortunately", she described an event involving one of the protagonists with a pronoun at the end of the sentence (e.g., "Unfortunately, the tooth paste spilled on him"). In the sentence, the subject NP referred to a key object, from which the relevant protagonist can be inferred. The final word was either "him" or "her", which completely disambiguated the relevant protagonist. The complete disambiguation in speech prevented unnatural focus and reliance on gestures. At the onset of the word "unfortunately" in the third sentence, the protagonists' names in black squares appeared (Picture 4 in Fig. 1). The black squares lasted until the end of the visual stimuli. Throughout the experiment, Gary was introduced first in each story.

As for gestures, in the first two sentences, gestures assigned the two protagonists to the actor's right and left frontal spaces with her right and left hand respectively when each protagonist was mentioned by the proper names in the first sentence. This was repeated for the pronouns in the second sentence (Pictures 2 and 3 in Fig. 1). The locations of the gestures for both protagonists were fixed; in other words, in all stories, Gary was always assigned on the right side and Betty was assigned on the left side from the participant's perspective. Thus, from the actor's perspective, Gary was assigned on her left-hand side and Betty was assigned on her right-hand side. The actor did not produce any gestures in the third sentence (Picture 4 in Fig. 1).

The congruent and incongruent conditions were created as follows. The locations of the protagonists' names in the block squares were on the same sides as the locations to which the actor gesturally assigned each protagonist in the congruent condition, but they were on the opposite sides in the incongruent condition. Thus, from the participant's perspective, Gary's name appeared on the right side and Betty's name appeared on the left side in the congruent condition, and the locations of those names were the other way around in the incongruent condition. As there is an equal number of congruent and incongruent trials, the cohesive gestures did not provide any useful information for the participant's task.

The experiment was conducted on a Dell laptop computer using E-prime software. The actor's speech was heard from Bose stereo headphones. The display was at a distance of 60 cm from the subjects. The left response key was the 'E' key, and the right response was the 'P' key.

1.1.3. Procedure

All participants were tested individually. The participants were instructed that they would see the protagonists' names in black squares appearing in the last part of each story and that the locations of the black squares for Gary and Betty might change from trial to trial. They were also asked to indicate which protagonist had an accident by pressing the key on the same side as the protagonist's name on the screen, as quickly and as accurately as they could. Each participant completed six practice trials and 40 experimental trials within an approximate duration of 20 min. The experimental trials consisted of two blocks, and the two blocks were presented without any break. The two blocks had the same 20 stories, and the presentation order of the stories was randomized within a block. In each block half of the stories were in the congruent condition, and the other half, the incongruent version. If a participant watched a particular story in the congruent condition in the first block, then she or he watched the story in the incongruent condition in the second block. Before each trial, an asterisk appeared as the fixation point in the center of the screen for 500 ms. Then a video stimulus was presented in full screen. The inter trial interval was 1000 ms. The trial ended upon any response given by the subject after the stimulus offset and lasted about 20 s. No feedback was given to the participants concerning the accuracy of their responses. Reaction time (RT) was recorded as the time between the onset of the subject NP in the third sentence and the moment a response key was pressed.

2. Results

We calculated the number of correct trials and the mean correct RTs (ms) for the two conditions (Table 1). We excluded the following trials from the reaction time analysis; 1) trials with responses within the first 100 ms after stimulus onset (as they were classified as error), 2) trials with responses that were more than three standard deviations from the mean of each participant (as we considered them to outliers), 3) trials with incorrect responses.

We conducted a paired samples *t*-test on the number of correct trials and the mean RTs. The result showed that the number of correct trials was not statistically different between the two conditions, *t* (19) = 0.89, n.s. However, the mean RTs in the congruent condition was significantly faster than that in the incongruent condition, *t* (19)

Table 1
The mean (SD) of correct trials and RTs in the congruent and incongruent conditions in Experiment 1.

Condition	Congruent	Incongruent
Number of correct trials	19.1 (1.6)	18.9 (1.8)
RTs (the entire exp.)	1741 (592)	1804 (545)

= 2.68, p = 0.02, d = 0.5. See Table 1 for descriptive statistics.

3. Discussion

When participants saw a target protagonist's name on the screen on the same side as the gesture for that protagonist had appeared in earlier sentences, their responses were faster than when they saw it on the opposite side to which the gesture appeared. Thus listeners can create a spatial representation of a protagonist based on a speaker's gestures, and more importantly, the representation automatically remains active in a subsequent sentence without a gesture. We claim that the activation was automatic because there were no task-specific strategic reasons to keep the gestured information activated in the third sentence without any gestures; that is, speech provided information relevant for selecting the correct response, while gestures did not.

Because gestures for the two protagonists always appeared on the same sides (Gary on the right side and Betty on the left side), it is not clear if spatial representations of the protagonists can be established within minimal discourse, in which each location is assigned to a protagonist only twice (once to establish a location for a protagonist, and then once to refer back). Thus, in the next experiment, we changed the names of protagonists and the locations of male and female protagonists in the gesture space for each trial to see whether a listener could create a spatial representation within each story.

4. Experiment 2

4.1. Method

4.1.1. Participants

Twenty native English speakers (15 female and 5 male) took part in this study. The participants' age ranged from 18 to 25 years (M = 19.65, SD = 1.42). They reported having normal or corrected-to-normal vision and audition.

4.1.2. Material and apparatus

The material and apparatus were identical to Experiment 1 except for the following three things. First, each story had a different pair of a male protagonist and a female protagonist (e.g., Tina and Colin). Second, unlike Experiment 1, the gesturally established locations for male and female protagonists were counter-balanced: the male on the right and the female on the left in a half of the stories and the female on the right and the male on the left in the other half. Third, the order in which the male and female protagonists were introduced in the discourse was counter balanced: the male and then the female in half of the stories, and the female and then the male in the other half. All the video stimuli and data can be downloaded from the following URL: https://osf.io/52qjc.

4.1.3. Procedure

The procedure was also the same as Experiment 1 except that there were only twenty trials as opposed to 40 trials in Experiment 1 where the stories were repeated twice. This is motivated by our desire to shorten the experiment and by a further analysis of Experiment 1 that indicated that the effect of congruency was weaker in the second 20 trials than in the first 20 trials. Ten stories were in the congruent condition and the other ten stories were in the incongruent condition. The presentation order of the 20 stories was randomized.

5. Results

We excluded the following trials from the reaction time analysis; 1) trials with responses within the first 100 ms after stimulus onset (as they were classified as error), 2) trials with responses that were more than three standard deviations from the mean of each participant, 3) trials with incorrect responses.

Table 2
The mean and standard deviation of correct trials and RTs in each condition in Experiment 2.

Condition	Congruent	Incongruent
Number of correct trials	9.9 (0.4)	9.5 (0.6)
RTs	2411 (564)	2615 (480)

We conducted a paired samples t-test on the number of correct trials and the mean RTs. The result showed that the number of correct trials in the congruent condition was statistically greater than that in the incongruent condition, t (19) = 2.33, p = 0.031, d = 0.79 (see Table 2). The mean RTs in the congruent condition was significantly faster than that in the incongruent condition, t (19) = 3.91, p < 0.001, d = 0.39 (see Table 2).

6. Discussion

We again found the compatibility effect between the gesturally assigned locations of the protagonists and the locations of the response keys. This effect was found even though the protagonists' names and the gesturally indicated locations for male and female protagonists varied for each story. Thus, listeners created a spatial representation of protagonists within minimal discourse for cohesive gestures (each protagonist was gesturally referred to only twice), and the representation automatically remained active in a subsequent sentence without a gesture.

The above results still leave an open question as to whether the spatial representation created by gestures facilitated or interfered with the key-press response, according to the conditions. Because the gesturally assigned locations of the protagonists and the locations of the response keys are consistent in the congruent condition, a gesture may facilitate the response. In contrast, because the mapping is inconsistent in the incongruent condition, a gesture may hinder the response. To reveal this, we added a speech-only baseline condition that consisted of audio and still image, and compared the performance in the speech-only condition with those in the congruent and the incongruent condition.

Previous studies have consistently found the interfering effect of stimulus-response incongruence on reaction time. However, they have not consistently shown the facilitating effect of stimulus-response congruence (see Hommel, 1993). Experiment 3 will reveal whether the phenomenon examined in our study is the same as in the previous studies. If gesture has the facilitating effect, the performance would be better in the congruent condition than the speech-only condition. If gesture has the interfering effect, the performance in the speech-only condition would be better than in the incongruent condition.

7. Experiment 3

7.1. Method

7.1.1. Participants

Thirty native English speakers (28 female and 2 male) participated in this study. The participants' age ranged from 18 to 21 years (M = 20.65, SD = 1.42). They reported having normal or corrected-to-normal vision and audition. Because we added the speech-only condition, we increased the number of participants in Experiment 3 to heighten the statistical power.

7.1.2. Material and apparatus

The material and apparatus were identical to Experiment 2 except for the following two things. First, we added the speech-only condition. In this condition, a still image of an actor who put her hands on her lap was displayed while the sound was playing. The onset time when the protagonists' names appeared on the screen for each story was the same as the onset time of the stories used in Experiments 1 and 2. Second, the total number of trials changed from 20 to 21 in order to make sure that each of the three conditions had equal number of trial; each condition has 7 trials. One new story was added. The other 20 stories and the pair of a male and a female protagonist for each story were identical to Experiment 2. All the video stimuli and data can be downloaded from the following URL: https://osf.io/52qjc.

7.1.3. Procedure

The procedure was also the same as Experiment 2 except that each participant completed 21 experimental trials with an approximate duration of 10 min. The presentation order of the 21 stories was randomized.

8. Results

We excluded the following trials from the reaction time analysis; 1) trials with responses within the first 100 ms after stimulus onset (as they were classified as error), 2) trials with responses that were more than three standard deviations from the mean of each participant, 3) trials with incorrect responses.

To examine differences of performance among three conditions, we conducted an analysis of repeated-measure ANOVA on the number of correct trials and the mean RTs with a correct choice with the three conditions as a within-subject factor (Table 3). A main effect of the condition was not found for the number of correct trials, $F(2, 58) = 1.17$, n.s., but found for the mean RT, $F(2, 58) = 10.01, p < 0.001, d = 0.45$. Tukey post hoc tests ($p < 0.05$) showed that the mean RTs in the incongruent condition was significantly slower than that in the congruent condition and the speech-only condition.

9. Discussion

Experiment 3 provided two findings. First, just as in Experiments 1 and 2, we again found the compatibility effect between the gesturally assigned locations of the protagonists and the locations of the response keys. Second, we found that the RTs in the incongruent condition was significantly slower than that in the speech-only condition, but we found no significant difference between the congruent condition and the speech-only condition. This indicates that the response key assignment that was incongruent with gesturally established spatial representations interfered with the comprehension of the subsequent sentence without a gesture. We found no evidence for facilitation effects on the congruent response key assignment.

10. General discussion

There were three main findings. First, the reaction time in the congruent condition was faster than that in the incongruent condition (Experiments 1–3). In these experiments, speech provided information useful for the task, but cohesive gestures did not, so there were no strategic reasons to keep the gesturally established spatial story representation active in the test sentence, which did not have any accompanying gesture. Nevertheless, the spatial story representation was automatically activated. These results indicate that listeners generated a spatial story representation based on the speaker's cohesive gestures, and the representation was *automatically* activated during a subsequent sentence without a gesture; that is, the *Active gestural discourse representation hypothesis* was supported. This result is not compatible with the *Semi-active gestural discourse representation hypothesis* (the semi-active gestural discourse representation requires additional cohesive gestures to become active again). The second main finding was that cohesive gestures can establish spatial representations within minimal discourse, in which each location is indicted only twice: once to establish a referent in a location, and then another time to refer back (Experiment 2). That is, cohesive gestures quickly establish spatial story representation. The third main finding is that the incongruent condition leads to worse performance than the baseline condition without any gestures in the entire discourse, and the congruent condition did not facilitate performance relative to the baseline (Experiment 3). As we discuss below, the lack of facilitation effects likely reflect the fact that participants did not see gestures as a valid cue because half of gestures were not useful for the task. Thus, the lack of facilitation effect does not entail that cohesive gestures generally do not facilitate processing of subsequent sentences.

Table 3
The mean and standard deviation of correct trials and RTs in each condition in Experiment 3.

Condition	Congruent	Incongruent	Speech-only
Number of correct trials	6.9 (0.4)	6.7 (0.7)	6.8 (0.4)
Mean RTs	2245 (499)	2481 (560)	2275 (579)

The current findings add to the literature in two important ways. First, this study showed that not only can listeners pick up spatial story representations established by cohesive gestures (Goodrich Smith & Hudson Kam, 2012; Sekine & Kita, 2015), but they can also maintain the representations in a subsequent sentence without further gestural cues. This is important, given that speakers do not produce gestures in every sentence they utter. The current result indicated that cohesive gestures can have more pervasive influence on discourse comprehension than previous studies would lead us to assume. Second, listeners *automatically* activated spatial representations encoded by cohesive gestures even when the task did not require the participants to do so. This finding supports the "integrated-systems hypothesis" (Kelly et al., 2010), positing that speech and gesture are tightly integrated and mutually and obligatorily interact in order to enhance language. At same time, our finding goes beyond the previous study by Hudson Kam and Goodrich Smith (2011), in which the task explicitly demanded participants to use spatial story representations established by cohesive gestures. In their study, maintaining the gesturally established representations was the only plausible strategy for participants to select their response in a forced choice task.

The automatic processing of gesturally encoded information has been seen not only for cohesive gestures in the current study but also for iconic and metaphoric gestures in previous studies. That is, people process gesturally encoded information even when the task does not require them to do so. For example, when the task was to make a judgement based only on speech stimuli, accompanying iconic gestures that were semantically congruent vs. incongruent influenced performance (e.g., Kelly et al., 2010) and ERPs (e.g., Kelly, Kravitz, & Hopkins, 2004). Neuro-imaging studies in which participants did not have any task also showed a similar effect of semantic congruency between speech and iconic gestures (e.g., ERP, Özyürek, Willems, Kita, & Hagoort, 2007; fMRI, Willems, Özyürek, & Hagoort, 2007). In an fMRI study on metaphoric gestures in which participants' task was simply to press a button when they saw a new visual stimulus, speech-gesture combination stimuli activated various areas of brain more strongly than speech-only and gesture-only stimuli (Straube, Green, Bromberg, & Kircher, 2011). Thus, when representational gestures (iconic, metaphoric, and deictic gestures; McNeill, 1992) accompany speech, gesturally encoded information seem to be automatically processed.

The current study found that gesturally created spatial story representation interfered with the key-press response in the incongruent condition, but it did not facilitate the response in the congruent condition. This finding is consistent with some previous studies on the stimulus-response compatibility, which have found the interfering effect of stimulus-response incongruence on reaction time, but not the facilitating effect of stimulus-response congruency on the reaction time (e.g., Craft & Simon, 1970; Gunter & Weinbrenner, 2017; Hommel, 1993; Kornblum et al., 1990; Weinbrenner, 2017).

A recent study on cohesive gestures suggested that facilitating effects in the congruent condition arise only when gestures provided valid cues for the task. Gunter and Weinbrenner (2017) set up the experimental situations where cohesive gestures did or did not disambiguate a target referent in a discourse, and examined participants' brain activities in the gesture-speech integration. Although they found no facilitating effect of gesture when participants were presented with three conditions (the congruent, the incongruent, and the speech-only condition), they found the facilitating effect when only two conditions (the congruent and the speech-only condition) were presented. The authors suggested that cohesive gestures facilitate processing only when participants considered **gestures to not be useful cues** for the task. In the experiment with three conditions, participants probably considered gestures as not useful for tracking references because gestures were useful 50% of the time and not useful 50% of the time. In contrast, in the experiment with only two conditions (without the incongruent condition), participants may have considered gestures to be useful because gesture were always useful. Thus, participants in our study were not likely to see gestures as useful cues to track references. This may be the reason why the current study did not find the facilitating effect in Experiment 3.

There are three questions for future studies. Firstly, it is not clear from our study how long the representation created by cohesive gestures lasts. The current study indicated that listeners created a spatial story representation based on the speaker's cohesive gestures, and the representation was activated during a subsequent sentence without a gesture. However, we do not know from our findings about how long the representation is available for the listener and whether he or she updates the representation when seeing new gestures that differently use locations from previous gestures. Secondly, it is not clear how post-stroke hold in our stimuli affected story representation. In our stimulus, the actor kept holding her hands in the air after each gesture stroke phase. Sekine and Kita (2015) pointed out that the held hand should help maintain the association between the location and the referent and contrast the two locations in gesture space with different meanings. Thus, the question is whether the listener can create the story representation without the post-stroke hold to the same degree. Thirdly, it is not clear whether the current findings can be observed across speakers of different languages. The effect of cohesive gestures on subsequent discourse comprehension has been shown in a limited range of populations: English-speaking adults in the UK by the current study; English-speaking adults in America by Goodrich Smith and Hudson Kam (2012) and Hudson Kam and Goodrich Smith (2011); Japanese-speaking children and adults in Japan by Sekine and Kita (2015). Thus, it is important to investigate whether this effect can be shown in speakers of other languages.

In conclusion, using the stimulus-response compatibility paradigm, we provided supporting evidence for the *Active gestural discourse representation hypothesis*, which states that the spatial story representation created by cohesive gestures automatically remains active throughout the discourse, even in sentences without any gestures, and influences discourse comprehension processes.

References

Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica*, *123*, 1–30.

Broaders, S. C., & Goldin-Meadow, S. (2010). Truth is at hand: How gesture adds information during investigative interviews. *Psychological Science*, *21*(5), 623–628.

Cassell, J., McNeill, D., & McCullough, K. E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and non-linguistic information. *Pragmatics and Cognition*, *7*(1), 1–33.

Chafe, W. (1987). Cognitive constraints on information flow. In R. S. Tomlin (Ed.), *Coherence and grounding in discourse* (pp. 21–51). Amsterdam: John Benjamins.

Cocks, N., Sautin, L., Kita, S., Morgan, G., & Zlotowitz, S. (2009). Gesture and speech integration: An exploratory study of a man with aphasia. *International Journal of Language and Communication Disorders*, *44*, 795–804.

Craft, J. L., & Simon, J. R. (1970). Processing symbolic information from a visual display: Interference from an irrelevant directional cue. *Journal of Experimental Psychology*, *83*, 415–420.

Goodrich Smith, W., & Hudson Kam, C. K. (2012). Knowing 'who she is' based on 'where she is': The effect of co-speech gesture on pronoun comprehension. *Language and Cognition*, *4-2*, 75–98.

Gullberg, M. (2006). Handling discourse: Gestures, reference, tracking, 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*. http://dx.doi.org/10.1162/jocn_a_01125.

Gunter, T. C., Weinbrenner, J. E. D., & Berndt, M. (2012). *Gesture related activity precedes the utterance of words in a covert fashion: Electrophysiological evidence from the lateralized readiness potential. Paper presented at the Neurobiology of Language Conference, Donostia-San Sebastian, Spain, 25–27 October*.

Gunter, T. C., Weinbrenner, J. E. D., & Holle, H. (2015). Inconsistent use of gesture space during abstract pointing impairs language comprehension. *Frontiers in Psychology*, *6*(80), 1–10.

Hommel, B. (1993). Inverting the Simon effect by intention: Determinants of direction and extent of effects of irrelevant spatial information. *Psychological Research Psychologische Forschung*, *55*, 270–279.

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.

Kelly, S. D. (2001). Broadening the units of analysis in communication: Speech and nonverbal behaviours in pragmatic comprehension. *Journal of Child Language*, *28*, 325–349.

Kelly, S. D., & Church, R. B. (1998). A comparison between children's and dults' ability to detect conceptual information conveyed through representational gestures. *Child Development*, *69*, 85–93.

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, *89*(1), 253–260.

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*, 260–267.

Kendon, A. (1994). Do gestures communicate: A review. *Research on Language and Social Interaction*, *27*, 175–200.

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility—A model and taxonomy. *Psychological Review*, *97*, 253–270.

McNeill, D. (1992). *Hand and mind*. Chicago: University of Chicago Press.

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

McNeill, D., & Levy, E. T. (1993). Cohesion and gesture. *Discourse Processes*, *16*(4), 363–386.

Morford, M., & Goldin-Meadow, S. (1992). Comprehension and production of gesture in combination with speech in one-word speakers. *Journal of Child Language*, *19*(3), 559–580.

Namy, L. L., Campbell, A. L., & Tomasello, M. (2004). The changing role of iconicity in non-verbal symbol learning: A U-shaped trajectory in the acquisition of arbitrary gestures. *Journal of Cognition and Development*, *5*(1), 37–57.

Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience*, *19*, 605–616. http://dx.doi.org/10.1162/jocn.2007.19.4.605.

Sekine, K., & Kita, S. (2015). Development of multimodal discourse comprehension: Cohesive use of space in gesture. *Language, Cognition, and Neuroscience*, *30*, 1245–1258.

Sekine, K., Sowden, H., & Kita, S. (2015). The development of the ability to semantically integrate information in speech and iconic gesture in comprehension. *Cognitive Science*, *39*, 1855–1880.

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.

Stec, K., & Huiskes, M. (2014). Co-constructing referential space in multimodal narratives. *Cognitive Semiotics*, *7*, 31–59.

Straube, B., Green, A., Bromberg, B., & Kircher, T. (2011). The differentiation of iconic and metaphoric gestures: Common and unique integration processes. *Human Brain Mapping*, *32*, 522–533. http://dx.doi.org/10.1002/hbm.21041.

Weinbrenner, J. E. D. (2017). *Abstract pointing: ERP and behavioral evidence for its role in reference tracking*. Unpublished doctoral dissertationLeipzig, Germany: University of Leipzig.

Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex*, *17*, 2322–2333.

Yoshioka, K. (2005). *Linguistic and gestural introduction and tracking of referents in L1 and L2 discourse*. Nijmegen: Radboud University.