Co-thought and co-speech gestures are generated by the same action generation process

Mingyuan Chu[1] & Sotaro Kita[2]

[1] Neurobiology of Language Department, Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH, Nijmegen, The Netherlands.

[2] Department of Psychology, University of Warwick, Coventry CV4 7AL, UK.

Correspondence concerning this article should be addressed to Mingyuan Chu, Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH, Nijmegen, The Netherlands. Email: mingyuan.chu@mpi.nl. Phone: 0031-24-3521307

Abstract

People spontaneously gesture when they speak (co-speech gesture) and when they solve problems silently (co-thought gesture). In this study, we first explored the relationship between these two types of gestures and found that individuals who produced co-thought gestures more frequently also produced co-speech gestures more frequently (Experiments 1 and 2). This suggests that the two types of gestures are generated from the same process. We then investigated whether both types of gestures can be generated from the representational use of the action generation process that also generates purposeful actions that have a direct physical impact on the world, such as manipulating an object or locomotion (the *action generation hypothesis*). To this end, we examined the effect of object affordances on the production of both types of gestures (Experiments 3 and 4). We found that individuals produced co-thought and co-speech gestures more often when the stimulus objects afforded action (objects with smooth surface) than when they did not (objects with spiky surface). These results support the *action generation hypothesis* for representational gestures. However, our findings are incompatible with the hypotheses that co-speech representational gestures are solely generated from the speech production process (the *speech production hypothesis*).

Key words: co-speech gesture, co-thought gesture, action generation, speech production, affordance

When speaking, people often spontaneously produce hand gestures (co-speech gestures). In this paper, we focus on gestures that depict actions, motions, and shapes, or gestures that point to a referent. These are called representational gestures (McNeill, 1992; Kita, 2000). Throughout this paper, we use the term *gesture* to refer specifically to representational gesture.

The production of co-speech gesture is tightly linked to speech production (McNeill, 1992). The way people verbally express a motion event affects the way they gesturally depict it (Kita & Özyürek, 2003). Prohibiting or allowing gesture can alter children's verbal explanations of Piagetian conservation tasks (Alibali & Kita, 2010), adults' choice of syntactic frames to express motion events (Mol & Kita, 2012), and their speech fluency in verbal descriptions with spatial contents (Rauscher, Krauss & Chen, 1996).

The tight link between co-speech gesture and speech has led some researchers to claim that co-speech gestures are solely generated from the speech production process. We call this class of hypotheses the *speech production hypothesis*. For example, the Growth Point Theory (McNeill, 1992, 2005, 2012) proposed that co-speech gesture and speech originate from the same representation, that is, from the same "growth point" (i.e., the minimal idea unit that combines images and words) during speaking. The Sketch Model (de Ruiter, 2000; de Ruiter & de Beer, 2013) proposed that co-speech gestures and speech are based on the same communicative intention. Co-speech gestures are generated in the conceptualization phase (Levelt, 1989) of speech production. During this phase, speakers realized their communicative intent by generating the propositional representation of speech contents and the imagistic representation of co-speech gesture

contents. Some versions of the Lexical Retrieval Hypothesis (Butterworth & Hadar, 1988) proposed that co-speech gestures are generated during the formulating phase (Levelt, 1989) of speech production. During this phase, speakers select lexical items from their mental lexicons, and co-speech gestures are generated from the semantic features of these lexical items (e.g., forms, directions, locations). Although these hypotheses disagree on which stage of the speech production process is responsible for generating co-speech gestures, they all hold that the generation of co-speech gestures is inseparable from the speech production process.

The close interaction between co-speech gestures and speech does not necessarily mean that co-speech gestures have to be solely generated from the speech production process. It has been repeatedly shown that gestures can express information that differs from or even contradicts the information expressed in speech (see Goldin-Meadow & Alibali, 2013, for a review). This discrepancy between the content of co-speech gestures and speech suggests that co-speech gestures, at least sometimes, may be generated from a process that is not part of the speech production process. Some researchers hypothesized that this process is the *action generation* process, which is responsible for generating purposeful actions that have a direct physical impact on the world, such as manipulating an object or locomotion (Hostetter & Alibali, 2008; Kita, 2000; Kita & Özyürek, 2003; Kita, 2014). Co-speech gestures are the representational use of such actions. We call this hypothesis the *action generation hypothesis*. According to this hypothesis, action-related representations are constantly activated in working memory when we speak. These representations automatically activate the action generation system, which generates

plans for appropriate actions. Co-speech gestures arise from these action plans. They are the representational use of actions because they do not interact with the physical world.

There has been some evidence for a close link between co-speech gestures and actions. For example, participants produce more co-speech gestures when they describe manipulable items (e.g. scissors) than when they describe non-manipulable items (e.g., fish; Pine, Gurney, & Fletcher, 2010; see also Feyereisen & Havard, 1999). In Hosttetter and Alibali (2010), participants were asked to describe patterns of dots and shapes either after they physically constructed the patterns with wooden sticks or after they viewed the patterns on a computer screen. Participants produce representational gestures at a higher rate when they have physically constructed the patterns than when they have only viewed the patterns. Results from these two studies are consistent with the *action generation hypothesis* because they show that action generation potential (Pine et al., 2010) or action generation experience (Hostetter & Alibali, 2010) can increase the production of co-speech gestures. However, in Pine et al. (2010) and Feyereisen and Havard (1999), the speech contents were different when participants described the manipulable and the non-manipulable items. Hosttetter and Alibali (2010) did not report whether speech production differed between the action and the viewing conditions. Therefore, it remains unclear whether the differences in gesture production between conditions were due to differences in the involvement of the action generation system or due to differences in speech content between conditions.

The *action generation hypothesis* is further supported by a study in which speech content was controlled (Cook & Tanenhaus, 2009). Participants were asked to solve a Tower of Hanoi problem either by moving real objects with their hands or by moving

objects on a computer screen with a mouse. They then described their solutions to a listener who would be solving the same problems later. Participants who solved the problem with real objects produced more gestures with grasping hand shapes and more gestures with higher and more curved trajectories than those who solved the computerized version of the problem. Importantly, the two groups used similar verbal descriptions. These results are consistent with the *action generation hypothesis* because specific action information was only reflected in speakers' co-speech gestures, but was not reflected in their concurrent speech. However, in this study participants could see their own actions during the problem-solving phase, so it is possible that the different forms of gestures in the two conditions may be due to different visuo-spatial (non-actional) representations rather than different actional representations. Thus, this study does not provide clear evidence that gestures' underlying representations are inherently actional.

To provide stronger support for the *action generation hypothesis*, we examined whether the frequency of co-speech gestures can be automatically affected by the properties of referent objects that are relevant to actions but not to speech. We manipulated the affordances of the stimulus object (mugs) by either presenting mugs with smooth surface or mugs with spiky surface. We elicited co-speech gestures by instructing participants to think aloud as they completed mental rotation of these mugs. We did not give participants any action task before this task. We then examined the effect of affordances (spiky vs. smooth) on participants' gesture rates.

Affordances are properties of an object that suggest how it can be acted upon (e.g., Norman, 1988). Evidence showed that affordances of objects such as their location, shape

and orientation lead to different reaching and grasping actions (e.g., Tucker & Ellis, 1998; Ellis & Tucker, 2000). According to the *action generation hypothesis*, participants should produce co-speech gestures less frequently when the stimulus object has a spiky surface than when it has a smooth surface, as objects with smooth surfaces afford action more strongly. In contrast, the *speech production hypothesis* predicts that the affordances of the stimulus objects should not affect the frequency of co-speech gestures because the affordances should not influence speech production.

In addition to co-speech gestures, people also spontaneously gesture when they solve problems during silent thinking (co-thought gestures). When people silently solve mental rotation problems in a non-communicative setting (e.g., when they are left alone and are recorded by a hidden camera), they spontaneously produced co-thought gestures, which simulated the manipulation or the rotation of the stimulus object (Chu and Kita, 2008, 2011). For example, they rotate their hands with the index finger and thumb opposed, as if to grasp and rotate the object. They also rotate their right index finger, as if to simulate the rotation of the object. Compared to co-speech gestures, co-thought gestures are much less well understood. The mechanism underlying the production of co-thought gestures remains largely unknown. The *action generation hypothesis* proposes that both co-speech and co-thought gestures are generated from the representational use of the action generation process. According to this hypothesis, the production of both co-speech and co-thought gestures should be affected by factors that influence the action generation process and there should be a systematic relationship between these two types of gestures.

There is evidence that co-thought and co-speech gestures share many properties, suggesting that the co-thought and co-speech gestures are generated by a common mechanism. People produce more co-speech gestures when speech production is more difficult than when it is less difficult (e.g., Kita & Davies, 2009, Melinger & Kita, 2007, Hostetter, Alibali & Kita, 2007, Rauscher, Krauss & Chen, 1996; Wesp, Hesse, Keutmann & Wheaton, 2001). They produce more co-thought gestures when silent problem-solving task is more difficult than when it is less difficult (e.g., Chu & Kita, 2011). Gesture rates dropped over the course of experiments, both when participants silently solved mental rotation problems (co-thought gestures) and when they verbally described their solutions to these problems (co-speech gesture; Chu & Kita, 2008).The representational content of both co-speech and co-thought gestures also changed from more object-anchored forms to less object-anchored forms over time (Chu & Kita, 2008). Suppressing both co-speech and co-thought gestures led to less frequent use of imagined physical movements of objects in the problem-solving strategy (Alibali, Spencer, Knox, & Kita, 2011). Although these parallel findings are compatible with the idea of a common mechanism for the production of co-speech and co-thought gestures, none of these studies directly examined the relationship between co-thought and co-speech gestures within the same individual. Furthermore, although co-thought gestures in these studies were produced in silence, people might have produced inner speech with their co-thought gestures. It is possible that the parallel findings between the co-thought and co-speech gestures were because that both types of gestures were produced with similar speech (covert and overt speech). To eliminate this alternative explanation, the present

study elicited co-thought gestures in a non-communicative task where speech production was suppressed by a simultaneous verbal shadowing task.

To summarize, the goal of the present study is to investigate the relationship of co-thought and co-speech gestures within the same individual and test the *action generation hypothesis* by examining whether both co-thought and co-speech gestures are affected by the affordances of the stimulus objects. In Experiment 1, we elicited co-thought gestures using a mental rotation task and co-speech gestures using a motion event description task. If both types of gestures are generated from the representational use of the action generation process, such as simulating the manipulation of stimulus objects or simulating the movements of stimulus objects, participants who produce co-thought gestures more frequently should also produce co-speech gestures more frequently than those who produce co-thought gestures less frequently. Experiment 2 sought to replicate the correlation found in Experiment 1 and rule out the possibility that the correlation was due to participants generating inner speech during co-thought gesture production. However, a positive correlation between co-thought and co-speech gestures can only be indirect support for the *action generation hypothesis*, because the positive correlation can be attributed to other non-action related factors as well. Experiment 3 and 4 sought to provide direct evidence for the *action generation hypothesis* by examining how action-related physical properties, namely the affordances of the stimulus objects affect the frequency of co-thought and co-speech gestures. We asked participants to solve a mental rotation task with a simultaneous verbal shadowing task to elicit co-thought gestures (Experiment 3) and to verbally explain their solution of a mental rotation task, to elicit co-speech gestures (Experiment 4). According to the *action generation hypothesis*,

participants should produce both co-thought and co-speech gestures more often when the stimulus objects afford action more strongly than when they are less likely to afford action.

Experiment 1

The main goal was to examine whether the rates of co-speech gestures correlated with the rates of co-thought gestures within the same individuals. If the two types of gestures are generated from the representational use of the action generation process, they should be positively correlated.

We also examined whether the rates of co-speech and co-thought gestures correlated with participants' rates of self-touches (e.g., scratching one's own body). This tested whether a positive correlation between the rates of co-thought and co-speech gestures was due to variations in the general tendency of moving one's hands while speaking or solving problems. In other words, people who are generally more likely to move their hands might produce both gestures and self-touch more often. If this were the case, people who produce co-thought and co-speech gestures very frequently should also produce self-touches very often. In contrast, if the positive correlation between the rates of co-thought and co-speech gestures was due to the representational use of the action generation process, there should not be any relationship between rates of self-touches and rates of the two types of gestures, because self-touches are not generated for representational purposes.

People spontaneously produce gestures not only when they talk to other people face-to-face, but also when they speak alone (Bavelas, Gerwing, Sutton, & Prevost, 2008; Cohen, 1977). Speakers gesture more often when speaking to a listener face-to-face than

when speaking alone (Cohen, 1977; Krauss, Dushay, Chen, & Rauscher, 1995; but see Bavelas & Healing, 2013 for a review). It is possible that gestures produced in these two situations may be generated from different mechanisms. For example, gestures produced in a face-to-face conversation may originate from communicative intent, whereas those produced alone may originate from non-communicative processes. Thus, the Sketch model (de Ruiter, 2000; de Ruiter & de Beer, 2013), which hypothesized that co-speech gestures originate from communicative intent, may predict that the frequency of co-thought gestures may correlate with the frequency of co-speech gestures produced in the speaking-alone situation, but not with the frequency of co-speech gestures produced in the face-to-face conversation. However, according to the *action generation hypothesis*, gestures are generated from the representational use of the action generation process regardless whether they are produced during silent problem-solving, in a face-to-face communication, or in a speak-alone situation. Thus, the rates of co-speech gesture produced in both situations should positively correlate with the rates of co-thought gestures.

Method

Participants. The participants were 41 native English speakers (37 female, mean age: 19 years old, age range: 18 - 28) from the University of Birmingham. All had normal or corrected-to-normal vision. They received course credits for their participation.

Mental rotation task. We elicited co-thought gestures by asking participant to solve a Shepard and Metzler (1971) type mental rotation task (see Figure 1 for an example and see supplemental materials for all stimuli). Each stimulus consisted of two three-dimensional objects presented at the top of the screen and one presented at the

bottom of the screen. The two upper objects were mirror images of each other on the vertical axis. They were always in the canonical position in the sense that their sides were parallel to the horizontal axis, the vertical axis, or the axis pointing to depth. The lower object was rotated from the upper left object in half of trials and from the upper right object in the other half of trials. The lower object was rotated by four angles (60 °, 120 °, 240 ° and 300 °) around the bisector that went through the object's center between the horizontal and vertical axis (XY axis), the horizontal and in-depth axis (XZ axis), and the vertical and in-depth axis (YZ axis). There were 24 experimental trials (left vs. right × 4 angles × 3 axes) and no practice trials. Stimuli were presented randomly.

Participants were asked to decide whether or not the lower object was rotated from the upper left or right object. In each trial, they first saw a white fixation cross in the center of the screen for 1000 ms and then the stimulus. As soon as they gave a response, the next trial started. They responded with two foot pedals, leaving their hands free for spontaneous gestures. They were told that accuracy was more important than speed so that spontaneous gestures were not suppressed due to time pressure. They were not told anything about gesture in the instructions. No feedback was given concerning the accuracy of their response. To maximally reduce the impact of the communicative environment, the experimenter left the room before the experiment started, and participants were left alone in the testing room. Their behaviour was recorded by a hidden camera (Sony DCR–HC19E PAL camcorder at 25 frames per second).

Motion event description task. We elicited co-speech gestures by asking participants to recount eight movie clips depicting movements of two geometric shapes (see supplemental materials). Each video clip was four seconds long. Each participant

described half of the clips in a face-to-face condition and the other half in a tape-recorder condition. In the face-to-face condition, the participant described the motion events to the experimenter sitting opposite the participant. The participants' behaviour was recorded by a video camera (Sony DCR–HC19E PAL camcorder at 25 frames per second), which was placed next to the experimenter and was visible to the participants. In the tape-recorder condition, participants were left alone in the room and described the motion events to a tape recorder. Their behaviour was video-recorded by a hidden camera (Sony DCR–HC19E PAL camcorder at 25 frames per second). There were no practice trials.

General Procedure. Participants were tested individually. They filled out the informed consent form, completed the mental rotation task, completed half of the motion event description task either in the face-to-face condition or in the tape-recorder condition, filled in personality questionnaires for about 30 minutes, and completed the other half of the motion event description task in the other condition. The questionnaire data were collected for a different study and are not reported in this paper. The order of the two conditions was counterbalanced across participants. After the participant completed the experiment, they were debriefed about the hidden video camera and its purpose and were given the opportunity to request erasing the recording. None of them reported awareness of the hidden camera. None of them requested to have their video data erased.

Gesture & self-touch coding. Gesture coding was carried out with video annotation software ELAN (European Distributed Corpora Project [EUDICO] Linguistic Annotator), developed by the Max Planck Institute for Psycholinguistics. Gestures were segmented according to the procedure described in Kita, Van Gijn, and Van der Hulst (1998). Each gesture was either categorized as a representational gesture or as a non-

representational gesture (on the basis of the classification system outlined in McNeill, 1992). Representational gestures are used to depict hand actions with objects, to represent physical properties or movements of objects, or to point to an object or a location. For example, in the mental rotation task, if a gesture was used to simulate manipulation of the stimulus object, to represent the rotation of the stimulus object, or to point to the stimulus object, it would be counted as a representational gesture. In the motion event description task, if a gesture was used to depict the shape of a stimulus object, to represent the manner and the path of a motion, or to point to an object or location, it would be counted as a representational gesture. Non-representational gestures included the following types of gestures: emblem or interactive gestures conveying conventionalized meanings, such as “maybe” (e.g., a flat hand with the palm down, wavering), “you know” (e.g., a flat hand with the palm up, possibly with a shoulder shrug); beat gestures were small, baton-like gestures produced along with the rhythm of speech to emphasize information; unclear gestures were gestures that could not be placed in any of the above categories.

Self-touches (also called "self-adaptors", Ekman & Friesen, 1969) or “body-focused movements” (Freedman, O'Hanoln, Oltman, & Witkin, 1972) were classified as hand movements that touched one’s own body or its adornments. Self-touches did not convey any information related to the speaking task or the mental rotation task.

To establish inter-coder reliability of gesture classification, a second independent coder classified the hand movements of eleven randomly selected participants (23% of all hand movements in the mental rotation task; 22% of all hand movements in the face-to-face condition of the motion event description task; 22% of all hand movements in the tape-recorder condition of the motion event description task). The two coders'

categorizations of representational, non-representational gestures, and self-touches matched on 98% of all hand movements (Cohen's $k = 0.94$, $p < .001$). To establish inter-coder reliability of gesture and self-touch identification, a third independent coder identified gestures and self-touches of the same eleven randomly selected participants. Among the gestures and self-touches that were identified by both coders, 97% of the original coder's gestures and 85% of the original coder's self-touches temporally overlapped with those identified by the third coder.

Results and Discussion

In the mental rotation task, participants produced 290 representational gestures, 34 non-representational gestures, and 501 self-touches. Twenty-five participants produced at least one representational gesture in the mental rotation task. In the motion event description task, they produced 756 representational gestures, 8 non-representational gestures, and 131 self-touches in the face-to-face condition; they produced 533 representational gestures, 8 non-representational gestures, and 160 self-touches in the tape-recorder condition. Fourty participants produced at least one representational gesture in the face-to-face condition and 36 participants produced at least one representational gesture in the tape-recorder condition.

We defined gesture rates as the number of gestures per minute. We used Spearman's rho for all correlation analyses because the distributions of co-thought gesture rates (Skewness = 2.85) and self-touch rates (Skewness = 1.50) in the mental rotation task were highly skewed (ps < .050).

We first examined the correlation between the rates of co-thought and co-speech gestures. To avoid influences of outliers, we excluded two participants whose gesture

rates were more than 2.5 standard deviations in the mental rotation task. No participants' gestures rates exceeded 2.5 standard deviations in the face-to-face or the tape-recorder condition of the motion event description task.

People who produced co-thought gestures more often also produced co-speech gestures more frequently, both in the face-to-face condition *rho* (37) = .49, *p* = .001 and in the tape-recorder condition *rho* (37) = .43, *p* = .009 (see Figure 2 for the scatter plots for the correlations).

We then examined the correlation between the rates of gestures and self-touches. We excluded two additional participants for this analysis. One participant's self-touch rates were more than 2.5 standard deviations in the mental rotation task and another participant's self-touch rates were more than 2.5 standard deviations in the face-to-face condition of the motion event description task. No participants' gestures rates exceeded 2.5 standard deviations in the tape-recorder condition of the motion event description task.

There was no correlation between the rates of self-touches in the mental rotation task and the rates of co-speech gestures in the motion event description task (in the face-to-face condition: *rho* (35) = -.10, *p* = .549; in the tape-recorder condition: *rho* (35) = -.07, *p* = .671). Similarly, there was no correlation between the rates of co-thought gesture in the mental rotation task and the rates of self-touches in the motion event description task (in the face-to-face condition: *rho* (35) = .16, *p* = .333; in the tape-recorder condition: *rho* (35) = -.07, *p* = .680). Furthermore, the rates of self-touch and gesture did not correlate in the mental rotation task (*rho* (35) = -.07, *p* = .664), in the face-to-face condition of the motion event description task (*rho* (35) = -.23, *p* = .166), or in the tape-recorder condition of the motion event description task (*rho* (35) = -.27, *p* = .112).

People who produced self-touch more often in the mental rotation task also produce self-touches more frequently in the motion event description task (in the face-to-face condition: *rho* (37) = .47, $p < .01$; in the tape-recorder condition: *rho* (37) = .30, $p = .061$).

The outlier exclusion was not crucial for the above results. Statistical significance or non-significance for all correlations for this experiment remained the same even if we included the outliers.

The positive correlation between the rates of co-thought and co-speech gestures is consistent with the idea that these two types of gestures are generated by a common mechanism. This common mechanism is unlikely to be a general tendency of moving one's hands while speaking or solving problems because the rate of self-touches did not correlate with the rates of gestures. This mechanism is also unlikely to be a part of the speech production process (e.g., Butterworth & Hadar, 1989; De Ruiter, 2000; McNeill, 1992) because co-thought gestures were produced in a non-speaking mental rotation task.

However, one could argue that the correlation between co-thought and co-speech gestures could be attributed to the possibility that co-thought gestures were triggered by inner speech when participants solved the mental rotation task. We aimed to rule out this possibility in Experiment 2.

Experiment 2

The goal was to replicate the positive correlation between the rates of co-thought and co-speech gestures, whilst also eliminating any possible inner speech when participants produced co-thought gestures. This was done by asking participants to count from one to five repeatedly while solving the mental rotation task. If both co-thought and

co-speech gestures are generated from the representational use of the action generation process, suppressing speech production in the mental rotation task should not affect the positive correlation between the rates of co-thought and co-speech gestures.

Method

Participants. The participants were 22 native English speakers (15 female, mean age: 21 years old, age range: 18 - 27) from the University of Birmingham. All had normal or corrected-to-normal vision. They were awarded course credits for their participation.

Mental rotation task. Participants completed the same mental rotation task as in Experiment 1, except that while solving the mental rotation problems, they was asked to simultaneously count from one to five aloud repeatedly according to the beeps at 0.4 second intervals heard through a headphone. Their behaviour was recorded by a hidden camera.

Motion event description task. The motion event description task was the same as the one used in Experiment 1. Participants' behaviour was recorded by a visible video camera in the face-to-face condition and by a hidden camera in the tape-recorder condition.

General procedure. The general procedure of Experiment 2 was the same as the one used in Experiment 1, except that participants were given the motion event description task immediately after the mental rotation task.

Coding. The gesture coding scheme were the same as the one used in Experiment 1. We did not code self-touches in this experiment.

To establish inter-coder reliability of gesture classification, a second independent coder classified the gestures of four randomly selected participants (23% of all hand movements in the face-to-face condition of the motion event description task; 29% of all hand movements in the tape-recorder condition of the motion event description task; 21% of all hand movements in the mental rotation task). The two coders' categorizations of representational and non-representational gestures matched on 98% of all gestures (Cohen's $k = 0.49$, $p < .001$). To establish inter-coder reliability of gesture identification, a third independent coder identified gestures of the same four randomly selected participants. Among the gestures that were identified by both coders, 92% of the original coder's gestures temporally overlapped with those identified by the third coder.

Results and Discussion

In the mental rotation task, participants produced 89 representational gestures, and 7 non-representational gestures. Fifteen participants produced at least one representational gesture in the mental rotation task. In the motion event description task, they produced 579 representational gestures and 4 non-representational gestures in the face-to-face condition and 418 representational gestures and 5 non-representational gestures in the tape-recorder condition. Twenty-one participants produced at least one representational gesture in the face-to-face condition and 20 participants produced at least one representational gesture in the tape-recorder condition. We calculated gesture rates by the number of gestures per minute.

Suppressing speech did not affect how often people produced co-thought gestures. The rates of co-thought gestures in Experiment 2 ($M = 1.03$, $SD = 1.12$) was not significantly different from the rates of co-thought gestures in Experiment 1 ($M = 1.23$,

SD = 2.11), *t* (61) = -.40, *p* = .688. This suggests that co-thought gestures are unlikely to be generated from speech production processes.

We used Spearman's rho for all correlation analyses because the distribution of co-thought gesture rates in the mental rotation task was highly skewed (Skewness = 1.51, *p*s < .050). To avoid influences of outliers, we excluded one participant whose gesture rate was more than 2.5 standard deviations in the mental rotation task. No participants' gestures rates exceeded 2.5 standard deviations in the face-to-face or the tape-recorder condition of the motion event description task. We replicated the findings of Experiment 1: People who produced co-thought gestures more often in the mental rotation task also produce co-speech gestures more frequently in the motion event description task (in the face-to-face condition *rho* (19) = .70, *p* < .001; in the tape-recorder condition: *rho* (19) = .53, *p* = .014, see Figure 3 for the scatter plots for the correlations)[1]. This indicates that the positive correlation between the rates of the two types of gestures was unlikely to reflect triggering of co-thought gestures by inner speech.

The outlier exclusion was not crucial for the above results. Statistical significance or non-significance for all correlations for this experiment remained the same even if we included the outliers.

The positive correlation between the rates of co-thought and co-speech gestures is consistent with the *action generation hypothesis* that both types of gesture are generated from the representational use of the action generation process. However, correlations are only indirect evidence for the *action generation hypothesis*, because it is possible that co-thought and co-speech gestures are generated by different processes, but that both processes are affected by common factors, such as the gesturer's spatial ability. Therefore,

in order to provide more direct evidence for the *action generation hypothesis*, we manipulated the factors that affect the action generation process and examined their effect on co-thought (Experiment 3) and co-speech gestures (Experiment 4).

Experiment 3

The goal was to provide direct evidence that *co-thought gestures* are generated from the representational use of the action generation process. Participants were asked to solve the same mental rotation task as used in Experiment 1, except that the stimulus objects were either mugs with spikes on their surface (less likely to be acted upon) or mugs with smooth surfaces (more likely to be acted upon; see Figure 4). If co-thought gestures are generated from the representational use of the action generation process, they should be sensitive to the affordances of the stimulus object. We predicted that the rates of co-thought gestures would decrease when there were spikes on the surface of the stimulus objects as people should be less likely to act on objects with spiky surfaces than objects with smooth surfaces.

Method

Participants. The participants were 24 native English speakers (19 female, mean age: 21 years old, age range: 18 - 25) from the University of Birmingham. All had normal or corrected-to-normal vision. They were awarded course credits for their participation.

Mental rotation task. The mental rotation task was similar to the one used in Experiment 2 except for the use of new stimuli, which consisted of two types of mugs (see Figure 4 for an example and see supplemental materials for all stimuli). In the spiky mug condition, 14 spikes were added to the original mug pictures (four spikes on the

handle, five spikes on each side of the mug). In the smooth condition, the mugs were presented with no spikes.

Participants were told that only one side of the mugs was painted in blue (note that the blue patch does not go all the way around the mugs in Figure 4). The handle of the upper left mug was on the left side of the blue surface, whereas the handle of the upper right mug was on the right side of the blue surface. Thus the two mugs on the upper screen were different from each other (i.e., the mirror image of each other).

There were 48 experimental trials presented randomly (spiky vs. smooth × left vs. right × 4 angles × 3 axes) and there were no practice trials. Each condition consisted of 24 trials. Participants solved mental rotation problems while counting simultaneously from one to five aloud repeatedly according to the beeps at 0.4 second intervals heard through a headphone. Their behaviour was recorded by a hidden video camera.

General procedure. The participants first completed the mental rotation task. They were told that surface differences (spikes vs. smooth) were irrelevant to the present study and should be ignored. The participants then rated the graspability of the smooth and the spiky mugs on a 1 to 5 scale (1 = least graspable; 5 = most graspable).

Gesture Coding. The gesture coding scheme was the same as the one used in Experiment 1. We did not code self-touches in this experiment.

To establish inter-coder reliability of gesture classification, a second independent coder classified the gestures of four randomly selected participants (28% of all hand movements). The two coders' categorization of representational and non-representational gestures matched 99% of all gestures (Cohen's $k = 0.93$, $p < .001$). To establish inter-coder reliability of gesture identification, a third independent coder identified gestures of

the same four randomly selected participants. Among the gestures that were identified by both coders, 92% of the original coder's gestures temporally overlapped with those identified by the third coder.

Results and Discussion

In total, participants produced 218 representational gestures and 32 non-representational gestures. Twelve participants produced at least one representational gesture. Smooth mugs (M = 4.46, SD = 0.83) were rated as more likely to be acted upon than spiky mugs (M = 1.96, SD = 1.08), t (23) = 10.07, $p < .001$, Cohen's d = 2.60.

Participants' gesture rates (i.e., number of gestures per minute) were higher in the smooth condition (M = 2.84, SD = 4.53) than in the spiky condition (M = 2.01, SD = 3.01), t (23) = 2.12, p = .045, Cohen's d = 0.22.

Can the gesture rate difference between the two conditions be attributed to the difference in difficulty of the two conditions? It has been shown that people gesture more often when people solve difficult problems than when they solve easy ones (e.g., Chu & Kita, 2011; Hostetter, Alibali & Kita, 2007; Kita & Davies, 2009; Melinger & Kita, 2007). However, the higher gesture rates in the smooth condition did not arise because people found smooth trials more difficult than spiky trials. On the contrary, perhaps because the spiky mugs were visually more complex than the smooth mugs were, participants found the spiky trials more difficult than the smooth trials. They needed longer RTs to solve each trial in the spiky condition (M = 3.35 seconds, SD = 1.47) than in the smooth condition (M = 2.90 seconds, SD = 0.98), t (23) = 3.05, p = .006, Cohen's d = 0.36[2]. Furthermore, the difference in gesture rates between the spiky and the smooth condition did not correlate with the RT differences between the two conditions, *rho* (22)

= .01, p = .979. Error rates in the smooth condition (M = 0.15, SD = 0.17) did not differ from error rates in the spiky condition (M = 0.15, SD = 0.17), t (23) = -.46, p = .649[3]. Thus, the gesture rate difference cannot be attributed to the difference in difficulty between the two conditions.

The analysis of individual differences in graspability ratings and gesture rates provided further evidence that objects affording an action elicited more gestures. Based on the differences in graspability ratings between the smooth and the spiky conditions (rating of the smooth mugs *minus* rating of the spiky mugs), we split participants at the median rating difference score (*median rating difference* = 3) into a high rating difference group (*mean rating difference* = 3.25,) and a low rating difference group (*mean rating difference* = 1.75,). Participants' gesture rates differences between the two conditions (gesture rates in the smooth condition – gesture rates in the spiky condition) were larger in the high rating difference group (M = 1.62, SD = 2.38) than in the low rating difference group (M = 0.05, SD = 0.88), t (22) = 2.15, p = .043, Cohen's d = 0.88.

Our results showed that co-thought gestures were affected by affordances of stimuli objects in the same way that actions would be affected. Participants' rating on affordances modulated the rates of their co-thought gestures. These results support the hypothesis that co-thought gestures are generated from the representational use of the action generation process.

Experiment 4

The goal was to provide direct evidence that *co-speech gestures* are generated from the representational use of the action generation process. Participants were asked to explain their solution to a similar mental rotation task as used in Experiment 3. If co-

speech gestures were generated from the representational use of the action generation process, people should produce co-speech gestures less often when they describe the rotation of spiky mugs than when they describe the rotation of smooth mugs.

Method

Participants. The participants were 23 native English speakers (22 female, mean age: 19 years old, age range: 18 - 21) from the University of Birmingham. All had normal or corrected-to-normal vision. They were awarded course credits for their participation.

Mental rotation description task. The stimuli were similar to those used in Experiment 3. Each stimulus display consisted of two same mugs at different orientations. The right mug was always in the canonical position. The left mug was rotated by four angles (60 °, 120 °, 240 ° and 300 °) around the Cartesian rotational axes (horizontal, vertical, and depth). At each angle for each axis, we presented either spiky mugs or smooth mugs (see Figure 5 for an example and see supplemental materials for all stimuli). There were 24 trials (spiky vs. smooth × 4 angles × 3 axes) presented randomly.

Participants were asked to describe how the left mug could be rotated to the position of the right one. They were asked to include the direction and angles of rotation in their description. They were told that surface differences (spikes vs. smooth) were irrelevant to the present study and should be ignored. They were also told that they were under no time pressure when solving the problems. The experimenter was seated to the left of the participants and pressed the space bar on the keyboard to start each trial. No feedback was given to the participants concerning the accuracy of their responses. Their behaviour was recorded by a visible video camera placed next to the experimenter.

General procedure. The general procedure was the same as that used in Experiment 3.

Gesture Coding. The gesture coding scheme was the same as that used in Experiment 1. We did not code self-touches in this experiment.

To establish inter-coder reliability of gesture classification, a second independent coder classified the gestures of three randomly selected participants (30% of all hand movements). The two coders' categorization of representational and non-representational gestures matched 98% of all gestures (Cohen's $k = 0.49$, $p < .001$). To establish inter-coder reliability of gesture identification, a third independent coder identified gestures of the same three randomly selected participants. Among the gestures that were identified by both coders, 95% of the original coder's gestures temporally overlapped with those identified by the third coder.

Results and Discussion

Participants produced overall 277 representational gestures and 8 non-representational gestures. Sixteen participants produced at least one representational gesture. Smooth mugs ($M = 4.43$, $SD = 0.79$) were rated as more likely to be acted upon than spiky mugs were ($M = 2.52$, $SD = 0.99$), $t\ (22) = 6.50$, $p < .001$, Cohen's $d = 2.13$.

Participants' gesture rates (i.e., number of gestures per minute) were higher in the smooth condition ($M = 3.74$, $SD = 5.17$) than in the spiky condition ($M = 2.81$, $SD = 4.35$), $t\ (22) = 3.82$, $p = .001$, Cohen's $d = 0.19$[4].

The higher gesture rates in the smooth condition did not arise because people found the smooth conditions more difficult than the spiky condition. The average number of words used in each trial, the average description duration in each trial, and the average

speech rates in each trial (i.e., number of words per minute) did not differ between the two conditions (see Table 1 for statistics)[5]. We did not measure description accuracy because participants were instructed to only estimate the rotation angle and thus the accuracy was not emphasized.

We also examined whether the difference in gesture rates between the smooth and the spiky conditions was due to differences in the content of the verbal descriptions in the two conditions. We categorized the words used in participants' description from all trials into either spatial words (e.g., left, clock-wise, thirty, degree), motoric words (e.g., turn, move, pull) or non-spatiomotoric words (e.g., you, will, mug; See Appendix for the exhaustive lists of the three types of words). We aggregated words from all participants and morphological variants (e.g., tilt vs. tilted vs. tilting).

The number of times each word was used in the two conditions was positively and very strongly correlated for all three types of words (spatial words: *rho* (56) = .90, $p < .001$; motoric words: *rho* (17) = .75, $p < .001$; non-spatiomotoric words: *rho* (83) = .79, $p < .001$; see Figure 6 for the scatter plots). So, for example, if the word 'left' was used very often and the word 'up' was used only a few times in the smooth condition, this was the case in the spiky condition as well.

The proportions of the spatial or motoric words out of all words did not differ between the two conditions. On average participants used 45% spatial words (*SD* = 0.10) and 10% motoric words (*SD* = 0.11) in the smooth condition and 46 % spatial words (*SD* = 0.05) and 10% motoric words (*SD* = 0.05) in the spiky condition (for spatial words: $t(22) = -0.77$, $p = .452$; for motoric words: $t(22) = 0.35$, $p = .728$).

We also carried out the same analysis of individual differences in graspability ratings and gesture rates as in Experiment 3. Based on the graspability rating differences between the smooth and the spiky conditions, we split participants at the median rating difference score (*median rating difference* = 2) into a high rating difference group (*mean rating difference* = 2.91, *n* = 11) and a low rating difference group (*mean rating difference* = 1, *n* = 12). Differences in participants' gesture rates in the two conditions were marginally significantly larger in the high rating difference group (*M* = 1.26, *SD* = 2.02) than in the low rating difference group (*M* = 0.02, *SD* = 0.89), *t* (21) = 1.93, *p* = .067. Thus, we found the same trend as in Experiment 3.

Our results show that co-speech gestures were affected by affordances of stimuli objects in the same way as actions would be affected. By contrast, the verbal explanations were not affected by the manipulation of affordances. In addition, there was a trend that participants' rating on affordance modulated the rates of their co-speech gestures. These results support the hypothesis that co-speech gestures are generated from the representational use of the action generation process.

Further Analyses on the effect of object affordance on the production of three subtypes of representational gestures

We further categorized the representational gestures in Experiments 3 and 4 into three subtypes, based on the widely used gesture classification system used in McNeill (1992). The first subtype is character viewpoint gesture, which was used to simulate hand actions upon the stimulus object. The crucial criterion for this type of gestures was that the participants had to make a grasping or holding hand shape, e.g., the index finger and the thumb were opposed or the two palms were opposed, as if grasping or holding the

stimulus object. The second subtype was observer viewpoint gesture, which was used to represent physical properties or movements of the stimulus object without any grasping or holding hand shape, e.g., a flat hand representing the stimulus object rotated around the wrist or a hand with the extended index finger drew a circle in the air. The third subtype was deictic gesture, which was used to point to the stimulus object without showing any physical properties or movements of the stimulus object.

According to the action generation hypothesis, all subtypes of representational gestures are generated from the representational use of the action generation process, such as manipulating an object or locomotion. Thus, all three types should be used less frequently in the spiky condition than in the smooth.

To establish inter-coder reliability of the classification of the three subtypes of representational gestures, a second independent coder classified the gestures from the same participants used for intercoder reliability check in Experiments 3 and 4. The two coders' categorization of character viewpoint, observer viewpoint and deictic gestures matched 96.79% of all gestures (Cohen's $k = 0.95$, $p < .001$) in Experiment 3 and matched 94.95% of all gestures (Cohen's $k = 0.90$, $p < .001$).

In Experiment 3, out of 218 representational gestures, there were 103 character viewpoint gestures (47.25%), 70 observer viewpoint gestures (32.11%), and 45 deictic gestures (20.64%). In Experiment 4, out of 277 representational gestures, there were 73 character viewpoint gestures (26.35%), 188 observer viewpoint gestures (67.87%), and 16 deictic gestures (5.78%).

We pooled the data from Experiments 3 and 4 to increase statistical power and examined the effect of stimulus affordance on the production of the three subtypes of

representational gestures. The rates of the three subtypes of representational gestures were respectively submitted to 2 ×2 ANOVA analysis with stimulus affordance (smooth vs. spiky mugs) as a within-participant factor and experiment (Experiment 3 vs. Experiment 4) as a between-participant factor. We only included participants who produced at least 1 representational gesture in these analyses (n = 12 in Experiment 3; n = 16 in Experiment 4).

For the rates of character view point gestures, there was a main effect of stimulus affordance (F (1, 26) = 7.32, p = .012, η^2= 0.22); that is, the rates of character viewpoint gestures were higher in the smooth condition (M = 1.97, SD = 3.05) than in the spiky condition (M = 1.32, SD = 2.23). There was no main effect of Experiment (F (1, 26) = 1.15, p = .294). There was no interaction between stimulus affordance and Experiment (F (1, 26) = 1.06, p = .312).

For the rates of observer view point gestures, there was a main effect of stimulus affordance (F (1, 26) = 13.12, p = .001, η^2= 0.34); that is, the rates of observer viewpoint gestures were higher in the smooth condition (M = 2.98, SD = 3.09) than in the spiky condition (M = 2.20, SD = 2.57). There was no main effect of Experiment (F (1, 26) = 2.35, p = .137). There was no interaction between stimulus affordance and Experiment (F (1, 26) = 1.08, p = .308).

For the rates of deictic gestures, there was no main effect of stimulus affordance (F (1, 26) = 0.42, p = .522). The rates of deictic gestures were only descriptively in the smooth condition (M = 0.59, SD = 1.36) than in the spiky condition (M = 0.51, SD = 0.94). There was no main effect of Experiment (F (1, 26) = 3.28, p = .082). There was no interaction between stimulus affordance and Experiment (F (1, 26) = 1.01, p = .325).

To summarize, the rates of character and observer viewpoint gestures were higher in the smooth condition than in the spiky condition. This was the case for both co-thought gestures elicited in Experiment 3 and co-speech gestures elicited in Experiment 4. The affordance of the stimulus objects did affect the production of the character and observer viewpoint gestures. Therefore, the character and observer viewpoint gestures in the present study are generated from the representational use of the action generation process. However, stimulus affordance did not affect the production of deictic gestures because the rates of deictic gestures were not significantly different between the smooth and the spiky conditions. Thus, it is unclear whether deictic gestures are generated from the representational use of the action generation process or from other processes (e.g., the speech production process).

General Discussion

Summary

The goal of the present study was to examine the relationship between co-speech and co-thought gestures, and to test the *action generation hypothesis*, which claims that both co-speech and co-thought gestures are generated from the representational use of the action generation process.

Experiment 1 and 2 showed that participants who produced co-thought gestures more frequently in a silent mental rotation task also produced co-speech gestures in a motion event description task more frequently. This positive correlation is unlikely to be due to individuals' general tendency to move their hands when talking or solving problems because their rates of self-touches did not correlate with gesture rates (Experiment 1). The positive correlation is unlikely to be attributed to the possibility that

co-thought gestures were triggered by inner speech, as the correlation was still observed when co-thought gestures were elicited during a non-communicative mental rotation task with a simultaneous verbal suppression task (Experiment 2).

The positive correlation between the co-thought and co-speech gestures is consistent with the *action generation hypothesis* that they are both generated from the representational use of the action generation process. The positive correlation between the two types of gestures is less consistent with the *speech production hypothesis* because co-thought gestures were unlikely to be generated from the speech production process.

Experiment 3 and 4 showed that co-thought and co-speech gestures are similar to real action (object manipulation) in terms of their response to object affordance. That is, people produced both co-thought and co-speech gestures less frequently when the stimulus objects were spiky mugs than when they were smooth mugs. The object affordances modulated production of both types of gesture in the same way they modulated actions: People tended not to act upon spiky objects. The lower gesture rates in the spiky condition than in the smooth condition cannot be attributed to differences in problem-solving difficulty across conditions (Experiment 3) or differences in speech content across conditions (Experiment 4). The idea that affordances influenced gesture rates was further supported by the correlational results that participants with a larger difference in graspability ratings (spiky = less graspable) showed a bigger affordance effect on gesture rates (spiky = lower gesture rates). These findings strongly suggest that co-thought and co-speech gestures are both generated from the representational use of the action generation process, which automatically takes into account whether or not a mug was spiky. Our results are in line with the finding that speakers encode action information

in their co-speech gestures, but not in their concurrent speech (Cook & Tanenhaus, 2009) and that speakers gesture more often when describing patterns they have physically constructed than when describing patterns they have only viewed (Hostetter & Alibali, 2010). Our results, however, go beyond previous findings because the affordance effect on the gesture rate in our experiments cannot be attributed to differences in prior visual experiences of gesturally depicted actions (Cook & Tanenhaus, 2009) or speech content (Hostetter & Alibali, 2010). More importantly, our results are not in accordance with the *speech production hypothesis* that co-speech gestures are generated from the speech production process.

Comparison of the action generation hypothesis with other hypotheses

The *action generation hypothesis* is in conflict with the speech production hypotheses that the generation of gesture is inseparable from the speech production processes. For example, according to the Growth Point Theory (McNeill, 1992, 2005), gesture and speech originate from a growth point that is an irreducible, minimal unit that combines imagery and linguistic categorical content. This hypothesis implies that the generation of gesture is inseparable from speech because gestures are not based solely on visuospatial imagery, but based on imagery that is, at the same time, a linguistic category, which will manifest itself as both a gesture and words. In addition, the Sketch Model hypothesis (De Ruiter, 2000) argues that gestures and speech originate from the same communicative intention generated for speaking; the Lexical Access Model (Butterworth & Hadar, 1988) proposes that gestures are produced from semantic representation of words retrieved for speech production. None of these hypotheses can explain why the affordance of stimulus objects should affect gesture production when the same affordance

does not affect speech content. Information about affordance was not part of the speakers' communicative intent because participants only describe the rotation direction and angle of the stimulus object without mentioning the spikes in their description.

The *action generation hypothesis* is compatible with the Information Packaging Hypothesis (Kita, 2000; also named as the Interface Model in Kita & Özyürek, 2003) and the gesture-as-simulated-action framework (Hostetter & Alibali, 2008). Both of them claim that gestures are "actions in the virtual environment" (Kita, 2000, p. 165) or "a natural expression of the simulated actions" (Hostetter & Alibali, 2008, p. 504). However, both hypotheses only address the origin of co-speech gestures and neither of them discusses the origin of co-thought gestures. The *action generation hypothesis* argues that both co-thought and co-speech gestures originated from the same mechanism: that is, the representational use of the action generation process. This claim has been supported by the findings that there is a positive correlation between the two types of gestures and both of them are affected by object affordances. This study for the first time provides direct empirical evidence about the relationship between co-speech and co-thought gestures.

It is worth pointing out that results of the current study indicates that gestures can be generated from the action generation process but they cannot tell us whether gesture can support the action generation process. To investigate the function of gesture, one needs to manipulate the availability of gesture and measure the effect of the action generation process. However, we will not be surprised if gesture facilitates the action-generation process or supports other non-linguistic cognitive processes (e.g., Alibali, Spencer, Knox & Kita, 2011; Chu & Kita, 2011; Pouw et al, 2014).

It is also worth mentioning that the action generation hypothesis does not deny the existence of close interaction between the gesture and speech production systems. It has been clearly shown that the interaction between gesture and speech can occur during both their planning and execution phases (e.g., Chu & Hagoort, 2014; Kita & Özyürek, 2003).

In addition, the action generation hypothesis and the speech production process may not be mutually exclusive. A gesture could be generated both from the action generation process and the speech generation process because the speech production process may also recruit the action-generation process. For example, when describing a cutting action, the speaker may perform a cutting gesture while saying the word "cut". In this case, both the gesture and the speech could originate from the cutting action generation process (Hostetter & Alibali, 2008). However, in Experiment 4, affordances of the mugs were irrelevant to speech production and were not mentioned in participants' speech at all. The effect of mug affordance on gesture rates was unlikely to be caused by speech-generation processes. Thus, results of Experiment 4 clearly argue against the hypotheses that co-speech gestures were solely generated from the speech production process.

Are all gestures generated from the representational use of the action generation process?

In the current study, representational gestures consisted of three subtypes of gestures, including character viewpoint gesture, observer viewpoint gesture and deictic gesture. The results showed that action related factors (e.g., affordance) affected not only those gestures that enact hand actions (character viewpoint gestures) but also those gestures that represent object motions and properties (observer viewpoint gestures). The

rates of deictic gestures did not differ between the smooth and the spiky conditions. This is perhaps due to the floor effect, in other words, because deictic gestures were infrequent in Experiments 3 and 4. Thus, the present study does not provide any direct evidence on whether deictic gestures are generated from the action generation process or the speech production process.

However, we speculate that deictic gestures might also be generated from the representational use of the action generation process because participants who produced deictic gestures more often also produced the other two types of representational gestures more frequently. The rates of deictic gestures were significantly positively correlated with the rates of non-deictic representational gestures (the combination of character viewpoint gestures and observer viewpoint gestures) in both Experiments 3 and 4 (Spearman's correlation, Experiment 3: *rho* (22) = .60, *p* = .002; Experiment 4: *rho* (21) = .48, *p* = .021). To draw firm conclusions on deictic gestures, future studies should use a task that can elicit more deictic gestures and examine the effect of stimulus affordance on the rates of deictic gestures.

Does our conclusion extend to gestures that metaphorically express abstract contents? People also gesture when talking about abstract concepts (McNeill, 1992). For example, when explaining the concept of conflict, speakers may move their hands toward each other as if the two hands, each holding an object, bang the two objects with each other (see Kita, Condappa, & Mohr, 2008 and Cienki & Müller, 2008 for more examples). Concrete and abstract concepts share common situational content, such as information about agents, objects, events (Barsalou & Wiemer-Hastings, 2005) and we understand abstract concepts in terms of image schemas based on concrete bodily experiences,

including actions (Johnson, 1987). Although the present study did not directly examine metaphorical gestures depicting abstract concepts, it is possible that these gestures are also generated from the representational use of the action generation process.

In addition to representational gestures, people also produce other types of gestures. These gestures include beat gestures (simple and rhythmic movements emphasizing the prosody or structure of speech without depicting semantic content related to speech), interactive gestures (movements used to manage the interaction between the speaker and the listener, such as an palm-up-open-hand gesture produced with “maybe” to show uncertainty) and emblem gestures (movements with specific meaning that are agreed within a community, such as an OK sign). These gestures are unlikely to be generated from the action generation processes. Further research needs to be done to study the origin of these gestures.

Limitations

One limitation of the current study is that the gestures observed in Experiments 3 and 4 were elicited by tasks with everyday manipulable objects (i.e., mugs) as stimulus objects. We used manipulable stimulus objects to maximize the chance of eliciting spontaneous gestures. However, one might argue that gestures elicited in Experiments 3 and 4 were more likely to be affected by action related factors, such as object affordances than objects that are not familiar (abstract 3-dimension objects) or not manipulable (houses, clouds). Future studies should explore to what extent present findings extend to other types of objects. Notwithstanding this limitation, the conclusion of the present study is clear: at least some co-speech and co-thought gestures are generated from the action generation process, but not from the speech generation process.

Furthermore, one might argue that the lower gesture rates in the spiky condition than in the smooth condition were due to differences in visual complexity between the two conditions rather than due to differences in object affordances. Although the current study cannot rule out this possibility, it seems unlikely. Previous evidence has shown that people gesture more when describing visually more complex diagrams than when describing simple diagrams (Kita & Davies, 2009). In the present study, participants gestured more in the smooth condition (visually less complex) than in the spiky condition (visually more complex). Furthermore, in the current study, individual differences in graspability ratings predicted individual differences in the effect size of the surface-type manipulation on gesture rates. This makes it less likely that visual complexity influenced gesture rates. It indicates that affordances had an impact on the gesture rate.

Conclusion

In sum, the present study provides both correlational and experimental evidence that, at least some co-speech as well as co-thought gestures are generated from the representational use of the action generation process. Whether or not to gesture is not only affected by what we are going to say but also by how our hands interact with the physical world.

Acknowledgments

We would like to thank Lucy Foulkes, Rachel Furness, Valentina Lee, and Zeshu Shao for their help with data collection, Paraskevi Argyriou for her help with reliability check of gesture coding, and Agnieszka Konopka and Josje Praamstra for their help with proof reading of this article.

References

Alibali, M. W., & Kita, S. (2010). Gesture highlights perceptually present information for speakers. *Gesture, 10,* 3-28.

Alibali, M. W., Spencer, R. C., Knox, L., & Kita, S. (2011). Spontaneous Gestures Influence Strategy Choices in Problem Solving. *Psychological Science, 22,* 1138-1144.

Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating abstract concepts. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thinking* (pp. 129-163). Cambridge: Cambridge University Press.

Bavelas, J. B., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language, 58,* 495–520.

Bavelas, J. B., & Healing, S. (2013). Reconciling the effects of mutual visibility on gesturing. A review. *Gesture, 13,* 63-92.

Butterworth, B., & Hadar, U. (1989). Gesture, speech, and computational stages: A reply to McNeill. *Psychological Review 96,* 168-174.

Chu, M., & Kita, S. (2008). Spontaneous gestures during mental rotation tasks: Insights into the microdevelopment of the motor strategy. *Journal of Experimental Psychology: General, 137,* 706-723.

Chu, M., & Kita, S. (2011). The nature of gestures' beneficial role in spatial problem solving. *Journal of Experimental Psychology: General, 140,* 102-116.

Chu, M., & Hagoort, P. (in press). Synchronization of speech and gesture: Evidence for interaction in action. *Journal of Experimental Psychology: General.*

Church, R. B., & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition, 23,* 43-71.

Cienki, A., & Müller, C. (Eds.). (2008). *Metaphor and gesture.* Amsterdam: John Benjamins.

Cohen, A. (1977). The communicative functions of hand illustrators. *Journal of Communication, 27,* 54–63.

Cook, S.W. & Tanenhaus, M. K. (2009). Embodied communication: Speakers' gestures affect listeners' actions. *Cognition, 113,* 98-104.

De Ruiter, J. P. A. (2000). The production of gesture and speech. In D. McNeill (Eds.), *Language and gesture* (pp. 284-311). Cambridge: Cambridge University Press.

De Ruiter, J.P. & de Beer, C. (2013). A critical evaluation of models of gesture and speech production for understanding gesture in aphasia. *Aphasiology, 27,* 1015-1030.

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica, 1,* 49–98.

Ellis, R. &Tucker,M. (2000) Micro-affordance: the potentiation of components of action by seen objects. *British Journal of Psychology, 91,* 451–471.

Feyereisen, P., & Havard, I. (1999). Mental imagery and production of hand gestures while speaking in younger and older adults. *Journal of Nonverbal Behavior, 23,* 153-171.

Freedman, N., O'Hanoln, J. O., Oltman, P., & Witkin, H. A. (1972). The imprint of psychological differentiation on kinetic behavior in varying communicative contexts. *Journal of Abnormal Psychology, 79,* 239-258.

Goldin-Meadow, S., & Alibali, M.W. (2013). Gestures role in speaking, learning, and creating language. *Annual Review of Psychology, 123,* 448-453.

Hostetter, A. B., Alibali, M. W., & Kita, S. (2007). I see it in my hands' eye: Representational gestures reflect conceptual demands. *Language and Cognitive Processes,* 22, 313–336.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin and Review, 15,* 495-514.

Hostetter, A. B. & Alibali, M. W. (2010). Language, gesture, action! A test of the Gesture as Simulated Action framework. *Journal of Memory and Language, 63,* 245-257.

Johnson, M. (1987). *The body in the mind: the bodily basis of meaning, imagination, and reason.* Chicago/London: The University of Chicago Press.

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Eds.), *Language and Gesture* (pp. 162-185). Cambridge, UK: Cambridge University Press.

Kita, S. (2014). Production of speech-accompanying gesture. In V. Ferreira, M. Goldrick & M. Miozzo (Eds.), *Oxford Handbook of Language Production* (pp.451-459). Oxford: Oxford University Press.

Kita, S., & Davies, T. S. (2009). Competing conceptual representations trigger co-speech representational gestures. *Language and Cognitive Processes, 24,* 761–775.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language, 48,* 16-32.

Kita, S., van Gijn, I., & van der Hulst, H. (1998). Movement Phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *Gesture and sign language in human-computer interaction, International Gesture Workshop Bielefeld, Germany, September 17-19, 1997, Proceedings. Lecture Notes in Artificial Intelligence, Volume 1371,* (pp. 23-35). Berlin: Springer-Verlag.

Krauss, R. M., Dushay, R. A., Chen, Y., & Rauscher, F. (1995). The communicative value of conversational hand gestures. *Journal of Experimental Social Psychology, 31,* 533-552.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

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

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

McNeill, D. (2012). *How Language Began: Gesture and Speech in Human Evolution.* Cambridge, UK: Cambridge University Press.

McNeill, D. & Duncan, S. D. (2000). Growth points in thinking-for-speaking. In D. McNeill (Eds.), *Language and Gesture,* (pp. 141-161). Cambridge, UK: Cambridge University Press.

Melinger, A. & Kita, S. (2007). Conceptual load triggers gesture production. *Language and Cognitive Processes, 22,* 473-500.

Mol, L. & Kita, S. (2012). Gesture structure affects syntactic structure in speech. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 761-766). Austin, TX: Cognitive Science Society.

Norman, D. A. (1988). *The Design of Everyday Things.* New York: Doubleday.

Pine, K. J., Gurney, D. J., & Fletcher, B. (2010). The semantic specificity hypothesis: When gestures do not depend upon the presence of a listener. *Journal of Nonverbal Behavior, 34,* 169-178.

Rauscher, F. B., Krauss, R. M., & Chen, Y. (1996). Gesture, speech and lexical access: The role of lexical movements in speech production. *Psychological Science, 7,* 226-231.

Reilly, D., & Neumann, D. L. (2013). Gender-Role Differences in Spatial Ability: A Meta-Analytic Review. *Sex Roles, 68,* 521-535.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171,* 701-703.

Tucker, M. & Ellis, R. (1998) On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 830–846.

Wagner, S. M., Nusbaum, H., & Goldin-Meadow, S. (2004). Probing the mental representation of gesture: Is handwaving spatial? *Journal of Memory and Language, 50,* 395–407.

Wesp, R., Hesse, J., Keutmann, D., & Wheaton, K. (2001). Gestures maintain spatial imagery. *American Journal of Psychology, 114,* 591 – 600.

Appendix

Non-spatiomotoric words	a, about, actually, again, and, as, be, bit, blue, but, by, can, case, certain, cup, dear, do, ehm, even, get, guess, handle, hard, have, how, I, instead, it, just, keep, know, leave, less, like, look, lot, many, matter, maybe, me, mean, more, motion, mug, nearly, need, no, not, of, oh, okay, or, possible, probably, quite, really, say, see, should, simple, so, some, something, sorry, sort, stages, than, that, the, them, then, thing, think, this, until, us, wait, way, well, which, whole, will, with, yeah, you
Spatial words	anti-clockwise, around, at, away, back, backwards, behind, bottom, clockwise, counter, degree, down, downwards, eight, eighty, fifteen, fifty, fifty-five, five, forty, forty-five, forward, from, hundred, in, left, leftwards, leftways, nine, ninety, ninety-five, on, one, over, right, rightwards, rightways, round, seventy, seventy-five, side, sixty, surface, ten, thirty, thirty-five, three, to, top, towards, twenty, twenty-five, two, up, upside, upwards, vertically, zero
Motoric words	bend, bring, come, facing, flip, going, hold, lift, move, pointing, pull, put, rotate, spin, take, tilt, tip, turn, twist

Footnotes

[1]The rates of co-thought and co-speech gestures were still positive correlated even after excluding the participants who did not produce any co-thought gestures in the mental rotation task. Statistical analyses are included in the supplemental materials.

[2] RTs in this experiment were calculated from the correct trials without any representational gesture. Including the trials with representational gestures did not change the results.

[3]Error rates in this experiment were calculated from the trials without any representational gesture. Including these trials did not change the result.

[4] The result remained the same when gesture rates were calculated by number of gestures per 100 words. Gesture rates were higher in the smooth condition (M = 5.43, SD = 7.42) than in the spiky condition (M = 4.31, SD = 6.68), t (22) = 3.41, $p < .01$, Cohen's d = 0.16.

[5]These three variables were calculated from the trials without any representational gesture. Including these trials did not change the results.