languages

MDPI

Article

Actions as a Basis for Online Embodied Concepts

Holly Keily

Department of Linguistics, University at Buffalo, Buffalo, NY 14260, USA; hmwillia@buffalo.edu

Received: 16 July 2018; Accepted: 26 February 2019; Published: 7 March 2019

Abstract: In co-speech gesture research, embodied cognition implies that concepts are associated with haptic and motor information that provides a framework for a gestural plan. When speakers access concepts, embodied action images are automatically activated. This study considers situations in which speakers need to create online concepts of events to investigate the aspect of the event that forms the basis of a new concept. Speakers watched short event video clips with familiar or unfamiliar attributes. They described those clips to partners who had to perform a matching task. Experimental results show that speakers gestured less and produced shorter gestures when relaying longer event descriptions. Speakers were more likely to produce gesture when some aspect of the event was unfamiliar, and they were most sensitive to the familiarity of the event's main action. Further, when speakers did gesture, they were most likely to gesture to represent the action of the event over the physical attributes of it (the instrument used to enact or the object acted upon). These findings suggest that in creating an embodied concept for something unfamiliar, the motion of the event acts as a basis for their online embodied representation of the concept.

Keywords: gesture; embodiment; cognition

1. Introduction

Spoken communication is fundamentally multimodal: our hands and words can encode propositions in tandem that our audience then decodes to derive meaning. In embodied cognition theories, a unified speech and gesture system provides information that makes up mental constructs. The constructs, then, necessarily consist of multidimensional information schema. In such embodied cognition theories, a word is indexed to perceptual experience, time-pressured, and embedded in the environment (Barsalou 1999, 2010; Clark 2008; Wilson 2002; Zwaan 2014). When we have new experiences, we create new conceptualizations. If gesture and speech are aspects of an integrated system, we can use gesture to examine how new concepts are formed by looking at how people represent the new information in both speech and gesture. The common conceptual root of both modalities can give us information on the process of embodied cognition, moving from perceiving to internalizing information. This research examines the possibility that when encountering novel events, the action attribute of the event is used as a ground for a new concept.

If a speaker describes an event they witnessed and produces gesture that represents only one attribute of that event (representing only the length of a fish when describing the event of catching said fish, for example), then that attribute has the highest activation, and is possibly the basis around which a mental representation of the event is built. The form and exact function of any individual gesture are probably unknowable, though it is likely that gestural form is related to the action of a simulation (Hostetter and Alibali 2008, p. 510). This study uses event descriptions to examine whether the action of an event possibly acts as the basis for an online embodied concept. If so, then the activation of the action should be higher than the activation of any other event attribute (such as the actor, the object acted upon, the instrument used, etc.), and speakers should produce more gesture relating to the action than any other part of the event.

When people produce speech and gesture together, their gestures tend to match what they say. People gesture in a way that elaborates on what they are saying. They move their hands as if holding a ball when talking about a ball. Their gestures can hold attention on the speaker, highlight or perform speech acts, and be integrated into speech at an information–structure level (Goodwin and Goodwin 1986; Heath 1992; Streek 1993). Researchers have established the tightly packaged meaning (Krauss and Hadar 1999; McNeill 1992), timing (Mayberry and Nicoladis 2000; McClave 1998), pragmatic functions (Goodwin and Goodwin 1986; Heath 1992; Kelly et al. 1999; Schegloff 1984; Streek 1993, 1994; Streek and Hartge 1992), and development (Butterworth 2003; Gullberg and Narasimhan 2010) of speech and gesture. This precise harmonization is used as support for the proposition that gestures and speech are coordinated at a high level, and are possibly two modes of expression from one system (Hostetter and Alibali 2008; Hostetter and Alibali 2010; McNeill 1992, 2005).

In the Sketch Model (de Ruiter 2000), as well as the Growth Point Theory (McNeill 1992, 2005) and the Gesture as Simulated Action (GSA) (Hostetter and Alibali 2008, 2010) models, the gestural plan for a concept is activated when the concept is activated. When people hear, read, or say a word, they access the concept: the word and the perceptual experience information are indexed to it. Action affordance studies are evidence that people develop action plans for handling a specific type of object when they hear the name of the object (Tucker and Ellis 2004). People have the same type of action plan activation when they see an image of an object, whether or not they intend to handle it (Fischer and Dahl 2007; Tucker and Ellis 1998). Additionally, when reading a word or seeing an image of an object, the different types of information inform different forms of gestural expression. Researchers have demonstrated that seeing an image of an object leads to more gesture specifically depicting the physical characteristics of the object, whereas reading the name of an object is correlated with relatively less gesture production but the produced gestures are focused on the utility of the object (Masson-Carro et al. 2017). Furthermore, objects with high affordances are associated with more representational gestures and with gestures based on how the object is manipulated in the real world (Masson-Carro et al. 2016). The perceptual/experiential information indexed with words is the basis for a gestural plan for those words and ultimately can be expressed in gesture.

For most speakers, sensorimotor experiences comprise a large part of the perceptual experiences people use in creating meaning (Zwaan 2014). Gestures are in theory either based on physical experience (action images) or a mental simulation of an experience (visual images). Action images come directly from physical experience, actual movements that are associated with concepts. They are mental representations ("images") characterized as motor plans for completing an action, based on experience. For example, speakers should have an action image for the experience of tying a shoelace. This image comes from the experience of having tied a shoelace previously. Visual images, while less durable than action images, in part explain why speakers can "draw" scenes they have witnessed: they are representing the mental exercise of manipulating a referent, but their image (and the resulting gesture) is not necessarily based in a pre-experienced motor plan (Hostetter and Alibali 2008). For example, a speaker might create a visual image when they are first learning to tie a shoelace. The image is a mental representation of the movements required to tie a shoelace, but it is not initially part of an established scheme.

In some instances, the stored embodied information in the action or visual image that is activated by a concept is sufficient to prompt a speaker to gesture. The GSA model hypothesizes that characteristics of the conversation, mental and physical abilities, as well as limitations, personal tendencies, and pragmatic considerations, can all conspire to affect a speaker's *gesture threshold*, an arbitrary and elastic level of activation that referent activation must surpass for a speaker to gesture (Hostetter and Alibali 2008, 2010). The gesture threshold is continuously adjusted to evolving dialogic situations and real-world considerations. For example, speakers holding large objects have heightened gesture thresholds due to that limitation. Speakers who tend to gesture frequently while they talk might have lowered gesture thresholds, as they are more disposed to gesturing. Sufficient activation can surpass that threshold and prompt speakers to gesture. When speakers do have sufficient activation to

prompt gesture, the gesture expresses spatially the [embodied] perceptual information associated with the aspect of the referent that is most activated.

Thus, constructed meanings are grounded in embodied action indexed to lexical labels. Action images (based in physical experience) from embodied cognition are most often the basis of action plans for speakers expressing a concept. If activation for the concept is high enough to surpass the speaker's gesture threshold, the speaker will gesture. Hence, the aspect of an utterance that is associated with gesture has the highest mental activation.

In the research below, speakers described events they had witnessed to a partner. The predictor variables hypothesized to influence speaker gesture production are (a) the type of event attribute (object, instrument, or action) and (b) its relative novelty for the speaker and hearer (familiar vs. unfamiliar). The dependent variables are the gesture production and gesture length. If actions are the basis for online embodied concepts, then action attributes will have higher activation than other event attributes and should be associated with more gestural representation. Some of the events were relatively routine such that an existing construct could be easily modified to represent the event. Other events were novel and required the speaker to create a concept of the event to relate to it later. The descriptions and corresponding gesture and gesture length show which attributes of the events had the most activation.

2. Materials and Methods

In the present study, I used a set of short video clips produced by the Max Planck Institute for Psycholinguistics known as the "Cut and break" video clips (Bohnemeyer et al. 2001). The clips each show one or two people with an object and an instrument. The person uses the instrument to *cut* or *break* the object in a variety of manners. This study considered a subset of the video library, that is, 18 target videos. Each video included one object and one instrument, and the objects and instruments for the target videos are listed in Table 1 below. There were two combinations, ax with carrot and knife with carrot, that were repeated twice: in one, the carrot was cut along the bias, and in the other, it was cut across the bias.

Table 1. Objects and instruments used in target stimuli.

Object
carrot
cloth
rope
stick
Instrument
knife [1]
hammer
ax [1]
point [2]

[1] The combinations of *carrot* and *knife* or *ax* repeated with the manner of action altered. In one instance, the *carrot* was cut along the bias. In a second instance, the *carrot* was cut across the bias. [2] The *point* instrument is most often a chisel, held to obscure its form and show the viewer only that the action was carried out with a large, solid, pointed object.

My sample consisted of English monolingual American or Canadian university students between 18 and 24 years old.[1] There were 15 self-identified women and 7 self-identified men who participated, all paired into dyads. One member of each dyad was assigned the role of *speaker* and the other

[1] Internal Review Board (IRB) approval for the study and subsequent elaboration was granted from the University at Buffalo Social and Behavioral Sciences IRB on 21 November 2014.

the role of *matcher*. The data came from a sample size of 11 speakers. Three groups consisted of one woman and one man each, two groups consisted of two men, and six groups consisted of two women. There were 845 total utterances produced by participants in the speaker role. A small proportion of these (37 utterances, 4.38%) was elaborative in response to matcher questions (that requested clarification about the speakers' descriptions of the witnessed events). Analyses were run with the entire set of 845 utterances and with the 37 elaborative responses removed (on 808 utterances). There was no meaningful difference between the larger (including elaborative utterances) and smaller (excluding elaborative utterances) data sets. For consistency in comparison, elaborative utterances were removed, and the analysis that follows considers 808 descriptive utterances.

The study is a modified referential communication task (Clark and Wilkes-Gibbs 1986). A participant (the *speaker*) watched a video of an event and described what they had witnessed to a partner (the *matcher*). After the initial description by the speaker, the matcher was presented with a matrix of images, one corresponding to the event. At this time, matchers were able to ask clarifying questions of speakers. The matcher selected the representational image most closely resembling the event. Each dyad completed the task for the target videos and ten supplementary videos. The supplementary videos were distractor items taken from the same library as the target videos. They were "cutting" and "breaking" events, but used non-matched objects, instruments, or actions.

I transcribed and coded for gesture each speaker's video description. The gestures were identified and coded based on a protocol adapted from the University of Chicago's McNeill Lab (McNeill and Duncan 2009). Movement was identified as meaningful (having some intentional form, aligning with speech, or replacing speech) or non-meaningful (e.g., self-touch or beat gestures). Abrupt changes in trajectory, speed, handshape, or interruption by a pause were indications of a separate gesture. For example, a hand with fingers closed toward the palm that moved from the speaker's shoulder height down to their midsection and back up with constant speed and with no other change in direction, as in a pantomime of using a hammer, was coded as a single gesture. If the speaker halted movement at the apex of the gesture path, then restarted by moving their hand up again and back down, that was identified as a second gesture. Gestures were associated with the meaningful attribute they represented: action, instrument, object, or other backgrounded information (including information about the actor and the setting).

Gesture production was coded as binary, with 0 = no gesture produced, 1 = gesture produced. The onset of a gesture was identified as continuous movement away from a speaker's resting position. All movement was annotated as a potential gesture phrase. After an initial annotation, identified movements were evaluated to remove beat gestures, self-touching, and stalled or otherwise unrealized gesture phrases. The remaining gesture annotations used in analysis included preparation, stroke, and retraction to resting position (Kendon 1980). After initial annotation was complete, an annotation tool, Embedded and Artificially Intelligent Vision Engineering (EAVISE), was used to check the onset and end of each gesture phrase. EAVISE identifies rest positions for hands and then measures when the hands move away from the rest positions with reliable accuracy (de Beugher et al. 2018). Gesture annotations were modified when necessary to align with the start and end times identified by EAVISE.

For attribute-specific gesture production, each meaningful (non-beat) gesture was coded as providing information about the object, instrument, action, or other. A single gesture could convey information about multiple attributes simultaneously. For example, a speaker might represent *cutting with a knife* by closing their fingers toward their palm with their thumb at the top of their fist and moving their hand downward and toward themselves, mimicking the action. This provides their audience with information about the action and instrument in a single gesture.

A ratio of attribute-specific gesture length and attribute-specific speech was used in the analyses. Hoetjes et al. (2015) evaluated both analyzing gesture rates per 100 words and per attribute. They concluded that the different rate analyses highlighted different aspects of the relationship between gesture and speech, with gesture per attribute calculations reflecting the early conceptual relationship between gesture and speech. In the present study, gesture per attribute was calculated

as milliseconds of gesture divided by the milliseconds of lexical description devoted to the same semantic attribute. In a few cases, some gestures clearly represented an event attribute, such as cutting, with no corresponding lexicalization. In these instances, no gesture length could be calculated so the gesture-to-speech ratio was set at one standard deviation higher than the maximum value, representing the relative significance of the amount of gesture produced compared to the amount of speech produced for the same attribute.

During the initial speaker descriptions, the matchers were not allowed to ask questions. After the initial description concluded, matchers could ask clarifying questions. Any speaker elaboration in response to clarifying questions was separated from the original description and transcribed and coded for gesture. These elaborations were found to be non-significant and were excluded from later analysis, as mentioned earlier. Description length was quantified as the length of time in milliseconds that the speaker used to describe the video clip, starting at the beginning of their first utterance and ending when they finished, before the matcher responded. Subsets of the total description were then identified as related to the attribute in focus for each portion of the description. Utterances that included information on the action, for example, were identified as action-focused. In the cases where a fluent utterance included information about more than one event attribute, the entirety of that utterance was coded as focusing on each attribute in its composite.

A separate online norming study elicited scores weighted by the plausibility of videos. Participants rated actions, instruments, and objects used in stimulus videos on a Likert-type scale assessing how familiar participants were with each event attribute. The scores were used to represent an average person's familiarity with attributes in each video for the target and supplementary videos. Familiarity was used to identify simulations where speakers would need to create online mental representations of the event. Participants in the norming survey watched the target videos. Participants watched a video then rated its attributes along a Likert scale according to how familiar they were with that aspect. For example, people were asked to rate their familiarity with the action shown in a clip (e.g., the *cutting* in a video of a person using a knife to cut a cloth). They repeated this for all 18 target and for 10 supplementary videos. Thirty-three respondents completed the norming task. All 33 respondents were required to have been in the United States while completing the task and to have self-identified as native English speakers. I have excluded data from respondents who rated every aspect as entirely familiar (7) or as entirely unfamiliar (0). Answers from 25 participants remained, each rating their relative familiarity with the action, instrument, and object for the videos. For the target videos, action ratings ranged from 2.13 to 6.46, instrument ratings ranged from 3.29 to 6.79, and object ratings ranged from 4.71 to 6.75.

3. Results

The main analyses tested the likelihood of a speaker producing gestures associated with different event attributes (object, instrument, action, or other, such as setting or actor) that varied in familiarity. I used generalized mixed-effects models with random effects for subjects and the video clips, with restricted maximum likelihood estimation (Baayen et al. 2008). Each model included interaction effects for description length, age, gender, and year in university. These demographic factors were included as nuisance variables to control for their possible influence on lexical and gesture production. The runtime of each stimulus video was recorded in milliseconds and used as a nuisance variable in analysis. I used Markov chain Monte Carlo simulations to estimate *p*-values.[2] Gestures were produced in 69.15% of the descriptions of the 18 target video clips.

---

[2] Models initially included linear and nonlinear transformations for variables, as there were multiple independent variables in consideration. Residuals from the linear transformations were distributed in a non-random pattern, whereas residuals from nonlinear transformations were not. Furthermore, the standard error of regression (S) for the linear transformation (65.50) was more substantial than S for the nonlinear transformations (13.92), indicating that the nonlinear transformations better fit the data. Thus, nonlinear transformations were used for analyses.

Of all the gestures produced, 14.62% provided information about objects, 21.83% provided information about instruments, and 63.55% provided information about actions (1.02% represented some other attribute, such as the physical appearance of the actor). Again, some gestures provided information about more than one event attribute. Speakers were four times (4.01) more likely to produce gesture with information about the action than the object, and almost three times (2.80) as likely to gesture with information about an action to an instrument.

3.1. Gesture Predicted by Attribute Type, Familiarity, and Description Length

The initial set of analyses were attribute-specific models, where the attribute in each stimulus clip (the specific object, instrument, or manner of action), the clip's attribute familiarity ratings, and the verbal description length were used as predictor variables with (a) gesture per attribute as a dependent variable (whether gesture was produced, coded as 1, or not produced, coded as 0), and (b) gesture length as a dependent variable (calculated as milliseconds of gesture divided by the milliseconds of lexical description devoted to the same semantic attribute). Demographic variables and video length were included as nuisance variables. Speakers and stimulus clips were treated as random effects.

In considering each object, instrument, and manner of action as distinct, the analysis targeted gestural and lexical realization of each event attribute as independent of one another. The resulting model examined whether an attribute was more likely to be represented gesturally based on attribute familiarity and the description time devoted to it, and if gesture length was similarly affected by attribute familiarity and description time.

Description length was a significant predictor of gesture production overall ($\beta = -3.44$, 95% CI = $[-6.17, -0.71]$, $p = 0.014$), and showed that speakers tended to gesture less when they had longer verbal descriptions. There was no main effect for attribute familiarity on gesture production.

Objects were not a significant predictor of object-specific gesture production ($p = 0.10$), perhaps reflecting the relative familiarity of objects used in the stimuli ($M = 6.34$ on a 7-point scale) compared to the instruments and manners of action. However, actions and instruments were also non-significant predictors of gesture per attribute ($p = 0.70$ and $p = 0.92$, respectively). Demographic variables and video length had no significant impact.

There was also no main effect for familiarity on gesture length. There was, however, an effect for a description length-by-attribute familiarity interaction on gesture length ($\beta = -0.038$, 95% CI = $[-0.042, -0.034]$, $p = 0.002$). The interaction was such that video clips with higher attribute familiarity ratings and short descriptions were associated with shorter gestures, whereas events with lower attribute familarity ratings and short descriptions were associated with longer gestures. There were no other significant effects on gesture length.

3.2. Gesture Predicted by Attribute Type, Attribute-Specific Focus and Familiarity

In a follow-up set of analyses, individual event attributes were again used as predictor variables along with the familiarity of each attribute and the attribute in focus. The attribute in focus was added as a predictor variable. Attribute in focus was coded as the amount of verbal description (in milliseconds) that conveyed information about different attributes of the event. As with gesture, the attribute in focus could contain information about multiple aspects of an event concurrently. The inclusion of attribute in focus as a predictor was meant to further understand the influence of description length on gesture production and how it might relate to the description length-by-attribute familiarity effect on gesture length.

These predictor variables were used to determine their effects on the dependent variables, namely, (a) number of gestures per attribute and (b) gesture length per attribute. Again, demographic variables and stimulus video length were included to control for their effects, and individual speakers and video clips identified as random variables.

There was a significant effect of attribute in focus on gesture production for action attributes: gesture production increased with increased proportionate description time devoted to the action of

the event (β = 3.44, 95% CI = [1.69, 5.27], $p = 0.008$). There was also an effect of action familiarity on gesture production, with more gesture for less familiar actions (β = −3.79, 95% CI = [−4.82, −2.34], $p = 0.02$).

The object as the attribute in focus was not significant in predicting gesture production. Similarly, the instrument in focus was not significant for gesture production. Furthermore, neither the familiarity of instruments nor objects had significant effects on gesture production. Interaction effects were not significant predictors of gesture production, nor were demographic variables or video length.

The attribute in focus variables did not prove useful predictors of gesture length. Furthermore, familiarity was not significant in predicting gesture length. The independent variable that most nearly met significance for predicting gesture length was action familiarity ($p = 0.053$). Interaction effects were also non-significant for gesture length.

4. Discussion

My results show that when describing events, speaker gesture production is sensitive to the length of the description and the relative familiarity of the action in the event. Shorter descriptions were accompanied by more gestures than longer descriptions, and short unfamiliar event descriptions were associated with longer gestures than were short familiar event descriptions. More specifically, events with unfamiliar actions were elaborated with gesture production more often than events with familiar actions. The familiarity of action affects the presence of gesture and is close to significance in predicting the length of gesture. Increased attention devoted to the action of an event also increases the likelihood of gesture production.

The general effect of familiarity on gesture production is driven mainly by the familiarity of the action. Instrument and object familiarity are independently insignificant, but their influences correspond to the direction of influence that action familiarity and overall familiarity have on gesture production. That is, speakers tend to gesture less when describing the familiar and more when they are describing the unfamiliar. Speakers form action images for events they witness. When the event is something for which there is an available schema, that schema is quickly and easily activated and used as a basis for the action image. When the event is highly unfamiliar, however, the action image is formed online and based on similar-enough existing representations. These action images must maintain high activation in order to remain accessible long enough for speakers to relay the information.

It is possible that the relationship between increased gesture, short description lengths, and decreased familiarity demonstrates the same effect, that gesture is being used as an aid to the speaker in advance or in place of speech. Previous studies have concluded that gesture is likely a strategy that compensates for the limits of working memory (de Ruiter 2000; Goodrich Smith and Kam 2012; Kendon 1972, 2004; Morsella and Krauss 2004; Wesp et al. 2001). Here speakers might be using gesture in part to help maintain a witnessed experience in working memory by physically simulating it while they search for appropriate labels. It is also possible that gesture is assisting speakers to counter the limitations of working memory by allowing them to remember from the environment, a proposition of some theories of embodied cognition (see Wilson 2002 for discussion).

Data for this study came from 11 speakers. The limited number of speakers affects the study's power and the generalizability of findings. Additionally, the sample was drawn from a fairly homogenous population of American university students. These individuals were all roughly the same age, and it is reasonable to assume that their life experiences, peer groups, and communication habits were all fairly similar. Effects of description length and familiarity found in the first analysis and effects of the action being in focus and action familiarity impacting the amount of gesture and gesture length were found. Further analysis with a larger, more varied population should be conducted to determine whether the effects hold.

The unfamiliar events in the stimulus clips were unfamiliar both to the speaker and the matcher. It is reasonable to expect that speakers recognized on some level that matchers would have no experience on which to base their interpretation of the description. It is possible that the speaker's

gesture production was a response to a communicative demand elevated by this awareness. Working in the GSA framework, a speaker's gesture threshold may be lowered with unfamiliar referents. This accounts for the experimental results that speakers gesture less when describing familiar compared to unfamiliar events. However, speakers were still more likely to produce gestures representing the action of events over the object or instrument in the event. Even if speakers have lowered gesture thresholds, the action of an event is the most highly activated.

The event aspect most often represented in gesture was also the attribute whose familiarity had the most impact on overall gesture production. When confronted with the unusual, speakers created new concepts. They built on existing constructs when possible, co-opting what appeared similar and embellishing to account for the new. Retaining this new concept to communicate it to someone necessitates a high activation of the construct. The observations from the experiment indicated that the event attribute with the most activation was the action. The action of an event must have the highest activation, as it is the attribute expressed gesturally. Even when gestures encode information about multiple aspects of the event, the action is often included. These findings support the Hostetter and Alibali proposition that gestural form is related to the action of a simulation (Hostetter and Alibali 2008, p. 510).

I propose further that the action of an event is a basis for its online embodied representation. Our embodied constructs for objects and instruments include sensorimotor information from handling them. They are elaborated with information about texture, color, hardness, weight, smell, and taste. This makes the construct specific, but also relatively cumbersome. Action constructs are also elaborate. They include information on the motor functions for performance, on speed, trajectory, and force. Yet action constructs are more generalizable as they are constrained by bodily limitations. There may be a hundred things similar to a zucchini (a summer squash, a cucumber, a plastic vegetable), but there is only one primary way to kick something. Therefore, when speakers create online representations of new events, they access the useful parts of existing constructs for actions, objects, and instruments. The action constructs, being largely generalizable, are easier to manipulate to the new situation and, by essentially being borrowings of existing concepts, are easy to maintain in working memory. With the action, speakers add the more elaborate conceptualizations for objects and instruments.

Funding: This material is based on work supported in part by funding from the Mark Diamond Research Fund under award FA-16-10.

Acknowledgments: I thank Jürgen Bohnemeyer for insight and expertise that greatly assisted the research. I also thank my colleague Saima Hafeez and the academic reviewers for comments that greatly improved the manuscript.

Conflicts of Interest: The author declares no conflict of interest.

References

Baayen, Rolf Harald, Douglas J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects. *Journal of Memory and Language* 59: 390–412. [CrossRef]

Barsalou, Lawrence W. 1999. Perceptual symbol systems. *Behavioral & Brain Sciences* 22: 577–660.

Barsalou, Lawrence W. 2010. Grounded Cognition: Past, Present, and Future. *Topics in Cognitive Science* 2: 716–24. [CrossRef] [PubMed]

Bohnemeyer, Jürgen, Melissa Bowerman, and Penelope Brown. 2001. Cut and break clips. In *Manual for the Field Season 2001*. Edited by Stephen C. Levinson and Nick J. Enfield. Nimegen: Max Planck Institute for Psycholinguistics, pp. 90–96.

Butterworth, George. 2003. Pointing is the royal road to language for babies. In *Pointing: Where Language, Culture, and Cognition Meet*. Edited by Sotaro Kita. Mahwah: Earlbaum Associates, pp. 9–33. ISBN 9780805840148.

Clark, Andy. 2008. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. New York: Oxford University Press, ISBN 9780195333213.

Clark, Herbert H., and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22: 1–39. [CrossRef]

de Beugher, Stijn, Geert Brône, and Toon Goedemé. 2018. A semi-automatic annotation tool for unobtrusive gesture analysis. *Lang Resources & Evaluation* 52: 433–60. [CrossRef]

de Ruiter, Jan-Pieter. 2000. The production of gesture and speech. In *Language and Gesture*. Edited by David McNeill. Cambridge: Cambridge University Press, pp. 284–311. ISBN 9780511620850.

Fischer, Martin H., and Christoph Dahl. 2007. The time course of visuo-motor affordances. *Experimental Brain Research* 176: 519–24. [CrossRef] [PubMed]

Goodrich Smith, Whitney, and Carla L. Hudson Kam. 2012. Knowing 'who she is' based on 'where she is': The effect of co-speech gesture on pronoun comprehension. *Language and Cognition* 4: 75–98. [CrossRef]

Goodwin, Charles, and Marjorie Harness Goodwin. 1986. Gesture and co-participation in the activity of searching for a word. *Semiotica* 62: 51–75.

Gullberg, Marianne, and Bhuvana Narasimhan. 2010. What gestures reveal about how semantic distinctions develop in Dutch children's placement verbs. *Cognitive Linguistics* 21: 239–62. [CrossRef]

Heath, Christian. 1992. Gesture's discreet tasks: Multiple relevancies in visual conduct and in the contextualization of Language. In *The Contextualization of Language*. Edited by Peter Auer and Aldo Di Luzio. Amsterdam: John Benjamins, pp. 101–27. ISBN 9781556192906.

Hoetjes, Marieke, Ruud Koolen, Martijn Goudbeek, Emiel Krahmer, and Marc Swerts. 2015. Reduction in gesture during the production of repeated references. *Journal of Memory and Language* 79–80: 1–17. [CrossRef]

Hostetter, Autumn B., and Martha W. Alibali. 2008. Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review* 15: 495–514. [CrossRef]

Hostetter, Autumn B., and Martha W. Alibali. 2010. Language, gesture, action! A test of the Gesture as Simulated Action framework. *Journal of Memory and Language* 63: 245–57. [CrossRef]

Kelly, Spencer D., Dale J. Barr, R. Breckenridge Church, and Katheryn Lynch. 1999. Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language* 40: 577–92. [CrossRef]

Kendon, Adam. 1972. Some relationships between body motion and speech. In *Studies in Dyadic Communication*. Edited by Aron Wolfe Siegman and Benjamin Pope. New York: Pergamon Press, pp. 177–210. ISBN 9780080158679.

Kendon, Adam. 1980. Gesticulation and speech: Two aspects of the process of utterance. In *Nonverbal Communication and Language*. Edited by Mary Ritchie Key. The Hague: Mouton Publishers, pp. 207–27. ISBN 9027978786.

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press, ISBN 9780511807572.

Krauss, Robert M., and Uri Hadar. 1999. The role of speech-related arm/hand gestures in word retrieval. In *Gesture, Speech, and Sign*. Edited by Lynn Messing and Ruth Campbell. Oxford: Oxford University Press, pp. 93–116. ISBN 9780198524519.

Masson-Carro, Ingrid, Martijn Goudbeek, and Emiel Krahmer. 2016. Can you handle this? The impact of object affordances on how co-speech gestures are produced. *Language, Cognition, and Neuroscience* 31: 430–40. [CrossRef] [PubMed]

Masson-Carro, Ingrid, Martijn Goudbeek, and Emiel Krahmer. 2017. How what we see and what we know influence iconic gesture production. *Journal of Nonverbal Behavior* 41: 367–94. [CrossRef] [PubMed]

Mayberry, Rachel I., and Elena Nicoladis. 2000. Gesture reflects language development: Evidence from bilingual children. *Current Directions in Psychological Science* 9: 192–96. [CrossRef]

McClave, Evelyn. 1998. Pitch and Manual Gestures. *Journal of Psycholinguistic Research* 27: 69–89. [CrossRef]

McNeill, David. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press, ISBN 9780226561349.

McNeill, David. 2005. *Gesture and Thought*. Chicago: University of Chicago Press, ISBN 9780226514635.

McNeill, David, and Susan Duncan. 2009. Annotative Practice, Revision of D. McNeill (2005). In *Gesture & Thought, Appendix*. Chicago: University of Chicago.

Morsella, Ezequiel, and Robert M. Krauss. 2004. The role of gestures in spatial working memory and speech. *The American Journal of Psychology* 117: 411–24. [CrossRef] [PubMed]

Schegloff, Emanuel. 1984. On some gestures' relation to talk. In *Structures of Social Action*. Edited by J. Maxwell Atkinson and John Heritage. Cambridge: Cambridge University Press, pp. 266–96. ISBN 9780521318624.

Streek, Jürgen. 1993. Gesture as communication I: Its coordination with gaze and speech. *Communication Monographs* 60: 275–99. [CrossRef]

Streek, Jürgen. 1994. Gesture as communication II: The audience as co-author. *Research on Language and Social Interaction* 27: 239–67. [CrossRef]

Streek, Jürgen, and Ulrike Hartge. 1992. Previews: Gestures at the transition place. In *The Contextualization of Language*. Edited by Peter Auer and Aldo Di Luzio. Amsterdam: Benjamins B.V., pp. 135–58. ISBN 9781556192906.

Tucker, Mike, and Rob Ellis. 1998. On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology: Human Perception & Performance* 24: 830–46. [CrossRef]

Tucker, Mike, and Rob Ellis. 2004. Action priming by briefly presented objects. *Acta Psychologica* 116: 185–203. [CrossRef] [PubMed]

Wesp, Richard, Jennifer Hesse, Donna Keutmann, and Karen Wheaton. 2001. Gestures maintain spatial imagery. *American Journal of Psychology* 114: 591–600. [CrossRef] [PubMed]

Wilson, Margaret. 2002. Six views of embodied cognition. *Psychonomic Bulletin & Review* 9: 625–36.

Zwaan, Rolf A. 2014. Embodiment and language comprehension: Reframing the discussion. *Trends in Cognitive Science* 18: 229–34. [CrossRef] [PubMed]