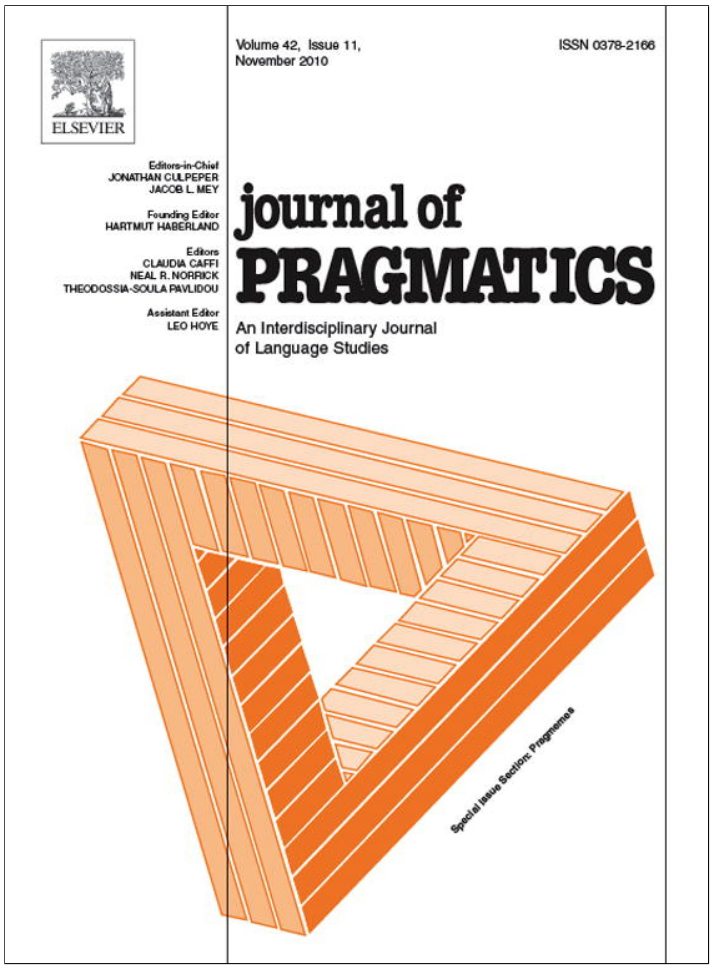

Journal of Pragmatics 42 (2010) 3130–3137

Contents lists available at ScienceDirect

Journal of Pragmatics

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

Effects of input modality on speech–gesture integration

Fey Parrill *, Jennifer Bullen, Huston Hoburg

Department of Cognitive Science, Case Western Reserve University, 10900 Euclid Ave, Cleveland, OH 44106, USA

ARTICLE INFO

Article history:
Received 7 May 2009
Received in revised form 12 April 2010
Accepted 17 April 2010

Keywords:
Gesture
Video

ABSTRACT

Do the gestures that accompany narrations produced after watching videos differ from those that accompany narrations produced after reading texts? Building on previous work by Hostetter and Hopkins, we ask whether participants gesture less after reading texts, whether participants produce different types of gestures after reading texts, whether the same motion event features (path and manner) are present in gesture in both cases, and whether gestural viewpoint (character or observer viewpoint) is impacted by input modality. We find that while participants produce longer narrations (and thus more gestures overall) after watching videos, gesture rate (number of gestures per word), gesture type, motion event feature encoding, and viewpoint are not affected by input modality. We comment on the implications of our findings for embodied theories of language.

1. Effects of input modality on speech–gesture integration

Research on co-speech gesture frequently uses video clips (cartoons, movies, and short action scenes) as stimuli in order to elicit narrative data. There are excellent reasons for using such stimuli. They evoke lively and gesture-rich data and are also non-linguistic, allowing one to test hypotheses about conceptual structure independent of linguistic structure. In addition, they provide an analyst with an objective basis for interpreting the meaning of gestures (McNeill, 1992). That is, when we see an upward moving gesture timed with an expression like *he went up* we tend to assume the gesture represents the character's upward motion, but our determination is based in part on the co-occurring speech. Knowing that upward motion occurs in a stimulus the narrator is describing provides a source of converging evidence for an interpretation of the gesture. While video stimuli are ideal for some purposes, they are also problematic. Some of the problems they present are methodological, while others are theoretical. In this study, we present a comparison between narrative data collected from participants who have seen video stimuli and from participants who have read text versions of those same stimuli (building on a similar study by Hostetter and Hopkins, 2002). We discuss the impact of stimulus modality (video versus text) on speech and gesture, and comment on the implications of our findings for embodied theories of language.

1.1. The use of video stimuli in research on co-speech gesture

A large number of studies on multimodal language use video stimuli. This tradition most likely arises from the fact that the first truly psychologically based (as opposed to sociologically based or anthropologically based) theory of co-speech gesture (McNeill, 1992) relied heavily on data collected using a cartoon stimulus (a now infamous cartoon called *Canary Row*, described in detail in McNeill, 1992). As a result, later work expanding on McNeill's initial findings tended to use the same stimulus or stimulus type (Alibali et al., 2001; Duncan, 2003, 2005; Kita, 2000; Kita and Özyürek, 2003; Kita et al., 2005; Levy and

* Corresponding author. Tel.: +1 216 368 2795; fax: +1 216 368 3821.
E-mail address: fey.parrill@case.edu (F. Parrill).

0378-2166/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.pragma.2010.04.023

McNeill, 1992; Parrill, 2008). Other researchers have used still images (often taken from cartoons or cartoon strips) in very similar ways (Beattie and Shovelton, 2002; Holler and Stevens, 2007; Wu and Coulson, 2005). Such stimuli share certain problems with video stimuli.

Obviously one's elicitation scenario should be tailored to one's research question. Indeed, there is an enormous body of research on gesture that does *not* use video stimuli. Some of this research comes from anthropological traditions, where data are collected naturalistically (Haviland, 2000; Kendon, 2004; Nuñez and Sweetser, 2006). In addition, many psychologists have collected experimentally controlled data using, for example, interactions surrounding props of various kinds (Emmorey et al., 2000; Furuyama, 2000; Gerwing and Bavelas, 2004; Goldin-Meadow et al., 2001; Ping and Goldin-Meadow, 2008) (For a nice collection of research coming from a variety of traditions, see Cienki and Müller, 2008.). Nevertheless, if a researcher's goal in studying multimodal language is to arrive at an understanding of how mental representations are encoded in speech and gesture, he or she must consider how the modality of input to short term memory plays a role in the encoding process. In short, while many different methods exist for collecting gesture data, there are both methodological and theoretical reasons for considering how the same basic content might be influenced by stimulus modality (video vs. text). We present a few of these reasons below.

1.2. Methodological problems with video stimuli

Video stimuli present methodological problems in that a researcher may or may not be able to generalize her findings to other types of data. This problem is shared by *all* elicitation scenarios, however, so will not be considered further. A more specific problem associated with video stimuli is that experimenters may need video stimuli with particular properties (e.g., certain visuo-spatial features). Of course, there are plenty of researchers for whom any video stimuli will do, so this problem is not universal. However, those who do need specific features in their stimuli must either find these stimuli, which can be time-consuming, or create their own, which can be costly. For researchers who do not need stimuli that are non-linguistic, knowing how well a text (which can be designed to fit one's needs exactly) can substitute for a video stimulus would be extremely helpful. Perhaps more important, researchers who actively want to use text stimuli to manipulate linguistic variables (e.g., verbal aspect, metaphorical language) need to know if the data they are collecting are different in any critical respect from those collected using videos. For example: Do participants gesture less? Do certain kinds of gestures not appear?

1.3. Theoretical problems with video stimuli

In addition to these methodological issues, it is important to understand how narrations based on video differ from those based on texts for theoretical reasons. First, texts are a central source of information for literate humans, and more data are needed that can inform us about gestures accompanying descriptions of texts. Second, mental representations based on video may be quite different from those based on text. Gestures display imagistic aspects of mental representations, so may reflect underlying differences in the how imagery is encoded. This issue is of central importance in the context of current *embodied* theories of language. Because such theories are gaining ground in linguistics, and because we feel our data provide a valuable source of support, we will elaborate on this point.

1.4. Gesture and embodied theories of language

The basic premise of embodied theories of language is that linguistic representations are not purely symbolic, but are instead modality specific. According to such theories, during language use we are *simulating*, or partially reconstructing, what is being talked about. Evidence is accumulating that motor and mental imagery are crucial parts of language. For example, Glenberg and Kaschak (2002), Kaschak and Glenberg (2000) have shown that performing an action incongruent with the meaning of a sentence slows response time in judging the sentence to grammatical. The same general effect can be shown for mental imagery tasks: participants are slower to respond to images that are incongruent with the meaning of a sentence (Zwaan, 1999; Zwaan et al., 2002). Processing sentences with implied vertical or horizontal semantic representations (e.g., *the ship sank* implies downward motion) also interferes with a spatial task that makes use of those same axes (Richardson et al., 2003). In summary, a number of studies suggest mental imagery and motor activation are a normal aspect of language comprehension. (See Bergen et al., 2007 for a review.) Thus far, however, researchers promoting embodied cognition have tended to ignore the fact that when we are producing language, we are frequently physically simulating what we are talking about through gesture. That is, motor and mental imagery emerge in gesture, and may be the product of simulation that takes place in the context of language use. Hostetter and Alibali (2008) make this point in a recent paper in which they integrate what is known about embodied language with what is known about gesture.

If simulation is a part of language use, and if gesture arises from simulation, does the modality of input have an effect on how an event is simulated? Video and text differ in at least two ways. First, a text will typically provide less imagistic detail than a video stimulus. Second, when describing a video, one is relying on visual imagery that has been encoded in short-term memory. While the resulting mental images will differ from the original visual images, they will retain a fair amount of detail from the original source, and will remain equivalent in a variety of ways (Kosslyn, 1994). The source of at least some of the content of a narration based on a video stimulus might therefore be described as speaker-*external*. Reading a text, on the other hand, requires one to generate visual and motor imagery based on mental representations associated with linguistic

symbols. While a text provides the linguistic symbols, recreating imagery and motion is a much more speaker-*internal* process. For these reasons, differences in simulation might be expected. While gesture is only an indirect measure of any (putative) process of simulation, it can nevertheless be very informative. In summary, there are a variety of reasons for wanting to know how speech and gesture are shaped by input.

2. Comparing video and text: targets of analysis

Previous research on co-speech gesture suggests that input modality might impact speech and gesture in a few specific ways. In particular, a previous study by Hostetter and Hopkins (2002) suggests that input modality will impact gesture frequency and gesture type. Hostetter and Hopkins asked participants to watch a video or to read a text description of the same video. They were primarily interested in comparing imagistic gestures (their *lexical movements*) with non-imagistic gestures (their *motor movements*—gestures which have no real semantic content). As a result, they divided gestures into these two categories. We attempt to replicate their findings, and also to extend their work by examining additional variables. We analyze gesture type using a finer-grained system. We also explore motion event features of gestures, and gestural viewpoint. Below we provide some explanation for why these variables are of interest.

2.1. Gesture frequency

As noted above, texts contain less imagistic detail than do video stimuli: this factor might also lead to a decrease in gesture frequency. Indeed, Hostetter and Hopkins found that participants who saw a video produced more representational gestures (their *lexical movements*) than did participants who read a text. We attempt replicate this finding with a larger dataset.

2.2. Gesture type

Gestures produced during narration vary in their semiotic properties (McNeill, 1992). Some gestures are primarily rhythmic (referred to as *beat* gestures). Others bear some resemblance to an action or entity described (*iconic* gestures). Some iconic gestures refer to abstract content (*metaphoric* gestures). Most gestures have some element of spatial or discourse deixis, but some gestures are primarily *deictic* (e.g., pointing gestures). While many gestures contain all these properties (rhythmicity, iconicity, metaphoricity, deixis), for the sake of analytic convenience, many researchers treat these semiotic dimensions as categories into which a gesture can be placed. We ask whether input modality has an impact on gesture type. In their 2002 study, Hostetter and Hopkins found that participants who saw a video produced the same number of beats (their *motor movements*) as participants who read a text, suggesting that some gesture types will be unaffected.

2.3. Motion event properties

The stimuli used in this study contain a great many motion events. As Talmy (1985) has shown, motion events can be decomposed into a series of features, including path (trajectory), and manner of motion (internal structure of an event). For example, an English sentence like *he hopped across the room* contains manner in the verb (*hop*), and path in the prepositional phrase. Gestures also encode information about these motion event properties. For example, the gesture shown in Fig. 1

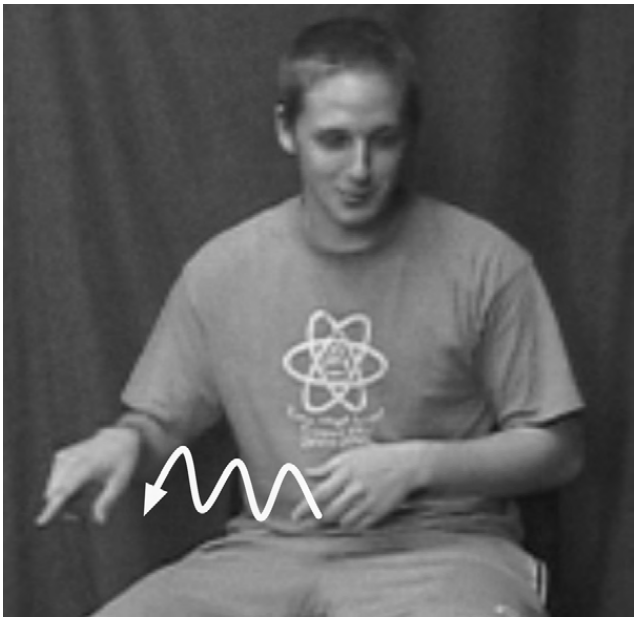

Fig. 1. Gesture encoding both path and manner.

Fig. 2. Character viewpoint gesture.

encodes both path and manner. A gesture that encoded only path might have a straight trajectory, while a gesture that encoded only manner might have an up and down component but no forward trajectory. Research has shown that the structure of the language one speaks influences which components are likely to appear in gesture (Kita and Özyürek, 2003; Kita et al., 2005; McNeill and Duncan, 2000).

A comparison between video and text offers interesting possibilities for exploring motion event encoding in gesture. As noted above, video stimuli provide an imagistic representation of a motion event, while texts provide a symbolic (linguistic) representation as input to a speaker's conceptualization. Because videos contain more detailed (and imagistic) information about a given motion event, more motion event features might appear in gestures produced after watching videos. On the other hand, if theories of embodied cognition are correct, the linguistic representations found in texts should (at least for these stimuli) prompt simulation. Simulation should result in the generation of an imagistic representation. As a result, gestures produced after reading texts might encode the same motion event information found in descriptions based on videos.

2.4. Viewpoint

Gestures can encode different points of view on the same basic event or scene. For example, a narrator may sometimes use her body as though she were a character. In Fig. 2, the narrator is describing a character hopping across a room. His arms are interpreted as the character's arms and his body as the character's body. The gestures produced in such situations have been called *character viewpoint* (C-VPT) gestures (McNeill, 1992). A narrator can also simply trace the path taken by the character with his finger, as shown in Fig. 1, where the narrator is describing the same hopping event. Such gestural depictions have been called *observer viewpoint* (O-VPT) gestures (McNeill, 1992). The narrator is depicting the action as though he were observing it from afar. Does input modality have any impact on viewpoint in gesture? MacWhinney (2005) proposes that stimuli rich in detail may induce participants to adopt a first-person perspective, which may be associated with character viewpoint gestures. Does the additional detail in a video stimulus make character viewpoint more frequent? On the other hand, if embodied theories of language are correct, a text stimulus should result in a rich simulation of the event described, thus character viewpoint gestures may be equally frequent.

3. Method

3.1. Participants and materials

Forty-six Case Western Reserve University students (26 women) participated in the study for payment. All were native speakers of English. Twenty-three participants (15 women) were randomly assigned to the *video* condition. These participants watched three thirty-second cartoon clips in random order. After watching each clip, participants described it to a friend who served as a listener for the study. To encourage narrators to describe the videos in detail, listeners took a comprehension quiz at the end of the study. Twenty-three participants (11 women) were randomly assigned to the *text*

condition. These participants read three text versions of the same cartoon stimuli, in random order, and described them to a friend. The text versions of the stimuli can be found in Appendix A, and were created on the basis of pilot narrations of the video stimuli, thus were designed to be representative of how narrators typically describe the events. Participants in both conditions were video/audio recorded as they described the stimuli to their partners.

3.2. Coding

All utterances produced by narrators were transcribed. Utterances were coded as *motion event* utterances if the event involved displacement of the agent or patient, or *non-motion* event utterances if there was no displacement. All gestures produced by participants were transcribed and sorted into the following categories by two independent coders: concrete iconic, metaphoric iconic, deictic, beat, adaptor (self-touching gestures: Ekman and Friesen, 1972), or emblem (conventional gestures such as thumbs up: (Ekman and Friesen, 1972). All gestures that occurred with a motion event utterance were coded for motion event information (path and manner). A gesture encoded path if the hand traced a trajectory. A gesture encoded manner if the internal structure of the action was represented in the gesture (usually with iterative motion). Some gestures encoded both simultaneously.

Concrete iconic gestures were also coded for viewpoint. Gestures in which the participant took on the role of the character, using his or her body as the character's body, were identified as C-VPT gestures. Gestures in which the participant took on the role of an observer, usually with hand or arm movements in the space in front of the body to reflect the character as a whole, were identified as O-VPT gestures. Very rarely, participants produced gestures that simultaneously reflected both viewpoints. These few instances were classified as *dual viewpoint*, but were extremely infrequent (1% in both conditions) and are not considered further. Speech was transcribed by a single coder. All gesture coding was carried out by two independent coders, for the entire dataset. Agreement was calculated using Cohen's kappa for gesture segmentation (that is, what counted as a gesture: $\kappa = .80$), gesture type ($\kappa = .72$), motion event information ($\kappa = .65$), and viewpoint ($\kappa = .86$). Landis and Koch (1977) suggest that any value of kappa above .61 indicates substantial inter-rater agreement. Disagreements were resolved through discussion to produce the final coded dataset.

4. Results

We first assessed the effects of input modality on gesture frequency. Table 1 shows the mean number of utterances and gestures produced by participants in the two conditions. Participants in the text condition produced marginally fewer gestures than participants in the video condition ($t_{44} = 1.62$, $p = .056$). (One-tailed p values will be reported because the predictions are directional.) This result does replicate the findings of Hostetter and Hopkins: when describing a text stimulus, participants gesture less. However, it is important to note that participants in the text condition produced significantly fewer utterances as well ($t_{44} = 2.13$, $p = .019$). That is, they also talked less. For this reason, it is important to look not just at gesture frequency (the overall number of gestures produced in each condition), but also at gesture rate (the number of gestures per word). The mean proportion of utterances accompanied by a gesture (the gesture *rate*) was not statistically different ($t_{44} = .33$, $p = .37$). Thus, it does not appear to be the case that reading a text makes one gesture less, but simply that reading a text makes one *talk* less. It should be noted that different researchers use different methods for calculating gesture rate (gestures per word, gestures per second). We also calculated gesture rate in terms of number of words to ensure comparability. More words were produced in the video condition ($t_{44} = 2.38$, $p = .01$), but gesture rate (number of gestures/number of words for each participant) was not significantly different ($t_{44} = .35$, $p = .36$).

4.1. Gesture type

Although participants produced gestures at the same rate across the two conditions, did they produce the same types of gestures? Table 2 shows the distribution of gesture types across conditions. (A small number of gestures that were unclassifiable have been excluded, thus the proportions do not add up to 1.) There was no significant difference across

Table 1
Mean number of utterances and gestures (standard deviation in parentheses).

	# Utterances	# Gestures
Video	53.5 (20.5)	38.6 (23)
Text	43.2 (10.72)	29.5 (14.2)

Table 2
Mean proportion of gesture types (standard deviation in parentheses).

	Concrete iconic	Metaphoric iconic	Beat	Deictic	Adaptor	Emblem
Video	.61 (.12)	.18 (.09)	.12 (.08)	.02 (.04)	.008 (.01)	.01 (.02)
Text	.57 (.20)	.11 (.07)	.18 (.18)	.02 (.03)	.01 (.02)	.01 (.02)

Table 3
Mean proportion of motion event features encoded in gesture (standard deviation in parentheses).

	Manner	Path	Path and manner
Video	.14 (.12)	.78 (.13)	.08 (.08)
Text	.14 (.16)	.74 (.21)	.06 (.08)

conditions for concrete iconic gestures ($t_{44} = .57$, $p = .28$), beat gestures ($t_{44} = 1.26$, $p = .10$), deictic gestures ($t_{44} = .03$, $p = .48$), adaptors ($t_{44} = .49$, $p = .31$), or emblems ($t_{44} = .12$, $p = .45$). However, participants in the video condition did produce significantly more metaphoric iconic gestures ($t_{44} = 2.78$, $p = .003$). We considered the possibility that because these participants produced longer narrations, they also produced more meta-narrative comments, accompanied by metaphoric gestures. However, there was no significant difference in the mean proportion of metanarrative utterances across conditions ($t_{44} = .004$, $p = .26$).

4.2. Motion event information

We also asked whether reading a text and watching a video have different effects on the motion event features that appear in gesture. Table 3 shows the mean proportion of each participant's gestures that encoded various motion event features. There was no significant difference across the two conditions in the production of gestures encoding manner of motion alone ($t_{44} = .19$, $p = .42$), path alone ($t_{44} = .83$, $p = .20$), or both path and manner simultaneously ($t_{44} = .63$, $p = .27$). Thus, input modality does not appear to impact the encoding of motion events for these stimuli.

4.3. Viewpoint

Finally, does reading a text have an effect on gestural viewpoint? For this analysis, only (concrete iconic) gestures expressing viewpoint were considered. The mean proportion of utterances accompanied by a character viewpoint gesture (with the remaining gestures encoding observer viewpoint) was .46 in the video condition (SD .15) and .54 in the text condition (SD .24). Although there was a tendency to produce more character viewpoint gestures after reading a text, this difference was not statistically significant ($t_{44} = 1.2$, $p = .11$). In short, input modality does not appear to affect viewpoint in gesture for these stimuli.

5. Discussion and conclusions

Watching a video and reading a text are two rather different experiences, and might be expected to give rise to quite different mental representations. Interestingly, the gestures produced in describing these two types of stimuli looked quite similar in our study. We suggested that texts contain less imagistic detail than do video stimuli, and that this factor might result in a reduced gesture rate. However, we found that reading a text resulted in a reduced length of narration, while gesture rates in the two conditions were not different. That is, when we control for the amount of speech participants produce, number of gestures is not affected.

Previous research indicated that more representational gestures might be expected with video stimuli. In our data however, most gesture types were unaffected, although small shifts across category did result in a larger number of metaphoric iconic gestures being produced in the video condition. Further research will be needed to determine whether this pattern is due to real differences in narrative structure across conditions. We also asked whether differences between video and text might have an effect on the encoding of motion event information (such as trajectory and manner of motion). Video stimuli provide an imagistic source for a speaker's conceptualization, and also contain more information about the motion events. Texts require a speaker to translate from a symbolic code to an imagistic representation, perhaps resulting in loss of detail about motion event properties. We saw no differences across our conditions. Nor did stimulus modality have an impact on viewpoint in gesture, with equal numbers of character and observer viewpoint gestures being produced regardless of condition. In other work, we have suggested that properties of an event itself may determine viewpoint in gesture in most cases (Parrill, 2010b), and to the extent that event properties are preserved across text and video, viewpoint will be unaffected.

Our results thus differ from those of Hostetter and Hopkins, who found that participants who saw a video produced more iconic gestures (their *lexical movements*) than did participants who read a text. (Our data do support the claim that the same number of beats will be produced.) We have presented a null result, which must always be interpreted with caution. The fact that we did not find a difference only means that this particular method failed to elicit one. We see two possible explanations for the difference we observe in gesture type when compared with Hostetter and Hopkins' findings. First, we attempted to account for the difference in overall narration length in assessing whether input modality impacts gesture production. That is, our analyses use gesture rate rather than overall number of gestures, since a larger number of gestures can be accounted for by a longer narration, irrespective of any experimental condition. Since Hostetter and Hopkins used overall frequency, we cannot compare our findings about gesture type directly to theirs. Second, in addition to the larger dataset, our study differs from Hostetter and Hopkins' in an important methodological way. In our study, participants gave their narrations to a friend,

while Hostetter and Hopkins asked participants to narrate to an experimenter. Participants produce fewer gestures and shorter narrations overall when narrating to an experimenter (Parrill, 2010a). This may occur either because they feel more inhibited, because they assume the information they are conveying is already known to their interlocutor, or both. While Hostetter and Hopkins did ask participants to "...pretend that the experimenter had never seen the cartoon they were describing" (p. 25), this request may not have been sufficient to mitigate the effect of narrating to an experimenter who was presumably a stranger. Our dataset is thus more naturalistic, but we leave open the possibility that the discrepancy between our findings and those of Hostetter and Hopkins are due to this methodological departure.

Although a null result should be treated cautiously, we have shown that video and text stimuli *can* evoke very similar gestural behaviors. We will end with some comments on the possible implications of this result. First, texts appear to serve as well as video stimuli in eliciting gesture, and the same types of gestures appear to be evoked by both sets of stimuli. Indeed, texts may reduce variation across narrations, which can be a desirable property. Second, imagery recalled from a video stimulus and imagery based on linguistic information may look very similar when schematized in gesture. This pattern offers some general support for theories of embodied cognition. Embodied theories of language claim that a linguistic prompt such as *she dashes up the stairs* should evoke motor simulation of the action of dashing as well as mental imagery of a figure's upward trajectory. The spontaneous production of co-speech gestures that encode visual imagery or motor representations can be seen as evidence for an underlying simulation. The same cannot be as easily said when the stimulus is a video. A person who has seen a visual image of the same event may simply be reenacting the event. That is, the trajectory and motor actions are already available externally, and can simply be imitated. However, we observed very consistent gestural behavior (in terms of rate, type, motion event properties, and viewpoint) across these two input types. This consistency suggests simulation plays a role in language production.

Acknowledgement

We thank Ben Bergen for helpful comments on an earlier version of this paper.

Appendix A

Text 1

There's a large bulldog being scolded by his owner for bringing too much junk into the house. While she is yelling at him, you can see he is hiding a kitten on his back. He slyly removes the kitten and hides it under a bowl on a shelf. Later, you see the kitten playing on a beach ball. The kitten is scrambling on top of the ball and it starts to roll towards the owner who is vacuuming in the next room. The ball bumps into the woman's leg and the bulldog looks alarmed.

Text 2

Bugs Bunny is the pitcher of a baseball game. After an extended windup, he pitches the ball, which is hit out of the park. Bugs runs out of the stadium to catch it. He takes a bus to the 'umpire state building' and then rides an elevator to the top. On the roof of the building, he attaches himself to a flagpole and pulls himself up. He still isn't high enough to catch it though, so he throws his glove up in the air. The glove catches the ball and then falls back on to Bugs Bunny's hand. The batter is out.

Text 3

Pepe Le Pew, the cartoon skunk, is cradling a black and white cat in his arms and kissing her. The cat scrambles out of Pepe's grasp and starts frantically scurrying around the house. Pepe Le Pew calmly hops along after her. She runs all around the room while he prances after her until she runs up the stairs and hides behind a door. She leans against the door and sighs with relief.

References

Alibali, Martha W., Heath, Dana C., Myers, Heather J., 2001. Effects of visibility between speaker and listener on gesture production: some gestures are meant to be seen. Journal of Memory and Language 44, 169–188.

Beattie, Geoffrey, Shovelton, Heather, 2002. An experimental investigation of the role of different types of iconic gesture in communication: a semantic feature approach. Gesture 1 (25), 129–149.

Bergen, B., Lindsay, S., Matlock, T., Narayanan, S., 2007. Spatial and linguistic aspects of visual imagery in sentence comprehension. Cognitive Science 31 (5), 733–764.

Cienki, Alan, Müller, Cornelia, 2008. Metaphor and Gesture. John Benjamins, Amsterdam.

Duncan, Susan D., 2003. Gesture in Language: Issues for Sign Language Research. Perspectives on Classifier Constructions in Sign Languages. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 259–268.

Duncan, Susan D., 2005. Gesture in signing: a case study from Taiwan Sign Language. Language and Linguistics 6 (2), 279–318.

Ekman, Paul, Friesen, Wallace, 1972. Hand movements. The Journal of Communication 22, 353–374.

Emmorey, Karen, Tversky, Barbara, Taylor, Holly A., 2000. Using space to describe space: perspective in speech, sign, and gesture. Spatial Cognition and Computation 26, 157–180.

Furuyama, Nobuhiro, 2000. Gestural Interaction between the Instructor and the Learner In Origami Instruction, Language and Gesture. Cambridge University Press, Cambridge, pp. 99–117.

Gerwing, Jennifer, Bavelas, Janet, 2004. Linguistic influences on gesture's form. Gesture 4 (2), 157–195.

Glenberg, A., Kaschak, M.P., 2002. Grounding language in action. Psychonomic Bulletin & Review 96, 558–565.

Goldin-Meadow, Susan, Nusbaum, Howard, Kelly, Spencer D., Wagner, Susan, 2001. Explaining math: gesturing lightens the load. Psychological Science 12 (6), 516–522.

Haviland, John B., 2000. Pointing, Gesture Spaces, and Mental Maps, Language and Gesture. Cambridge University Press, Cambridge. MA, pp. 13–46.

Holler, Judith, Stevens, Rachel, 2007. The effect of common ground on how speakers use gesture and speech to represent size information. Journal of Language and Social Psychology 26 (1), 4–27.
Hostetter, Autumn B., Alibali, Martha W., 2008. Visible embodiment: gesture as simulated action. Psychonomic Bulletin & Review 15 (3), 495–514.
Hostetter, Autumn B., Hopkins, William D., 2002. The effect of thought structure on the production of lexical movements. Brain and Language 82 (1), 22–29.
Kaschak, M.P., Glenberg, A.M., 2000. Constructing meaning: The role of affordances and grammatical constructions in sentence comprehension. Journal of Memory & Language 43 (3), 508–529.
Kendon, Adam, 2004. Gesture: Visible Action as Utterance. Cambridge University Press, Cambridge.
Kita, Sotaro, 2000. How representational gestures help speaking. In: McNeill, D. (Ed.), Language and Gesture. Cambridge University Press, Cambridge, MA, pp. 162–185.
Kita, Sotaro, Özyürek, Asli, 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 & Language 48 (1), 16–32.
Kita, Sotaro, Özyürek, Asli, Allen, Shanley, Furman, Reyhan, Brown, Amanda, 2005. How does linguistic framing of events influence co-speech gestures? Insights from cross-linguistic variations and similarities. Gesture 5 (1/2), 219–240.
Kosslyn, Stephen, 1994. Image and Brain: The Resolution of the Imagery Debate. The MIT Press, Cambridge, MA.
Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174.
Levy, Elena T., McNeill, David, 1992. Speech, gesture, and discourse. Discourse Processes 155, 277–301.
MacWhinney, Brian, 2005. The emergence of grammar from perspective taking. In: Pecher, D., Zwaan, R. (Eds.), Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking. Cambridge University Press, Cambridge, pp. 198–223.
McNeill, David, 1992. Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press, Chicago.
McNeill, David, Duncan, Susan D., 2000. Growth points in thinking-for-speaking. In: McNeill, D. (Ed.), Language and Gesture. Cambridge University Press, Cambridge, MA, pp. 141–161.
Nuñez, Rafael, Sweetser, Eve E., 2006. With the future behind them: convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construals of time. Cognitive Science 30 (5), 401–450.
Parrill, Fey, 2008. Subjects in the hands of speakers: an experimental study of syntactic subject and speech–gesture integration. Cognitive Linguistics 19 (2), 283–299.
Parrill, Fey, 2010a. The hands are part of the package: gesture, common ground, and information packaging. In: Rice, S., Newman, J. (Eds.), Empirical and Experimental Methods in Cognitive/Functional Research. CSLI Publications, Stanford, pp. 285–302.
Parrill, Fey, 2010b. Viewpoint in speech–gesture integration: linguistic structure, discourse structure, and event structure. Language and Cognitive Processes 25 (5), 650–668.
Ping, Raedy, Goldin-Meadow, Susan, 2008. Hands in the air: using ungrounded iconic gestures to teach children conservation of quantity. Developmental Psychology 44 (5), 1277–1287.
Richardson, Daniel C., Spivey, Michael J., Barsalou, Larry W., McRae, Ken, 2003. Spatial representations activated during real-time comprehension of verbs. Cognitive Science 27 (6), 767–780.
Talmy, Leonard, 1985. Lexicalization patterns: semantic structure in lexical forms. In: Shopen, T. (Ed.), Language Typology and Syntactic Description, vol. 3. Cambridge University Press, Cambridge, pp. 57–149.
Wu, Ying C., Coulson, Seana, 2005. Meaningful gestures: electrophysiological indices of iconic gesture comprehension. Psychophysiology 425, 654–667.
Zwaan, Rolf A., 1999. Embodied cognition, perceptual symbols, and situation models. Discourse Processes 28 (16), 81–88.
Zwaan, Rolf A., Stanfield, Robert A., Yaxley, Richard H., 2002. Language comprehenders mentally represent the shapes of objects. Psychological Science 136, 168–171.