Published in Cognitive Science 34, issue 8, 1517-1529, 2010
which should be used for any reference to this work

Effects of Ambiguous Gestures and Language on the Time Course of Reference Resolution

Max M. Louwerse,[a] Adrian Bangerter[b]

[a]*Department of Psychology, University of Memphis*
[b]*Department of Applied Psychology, University of Neuchâtel*

Abstract

Two eye-tracking experiments investigated how and when pointing gestures and location descriptions affect target identification. The experiments investigated the effect of gestures and referring expressions on the time course of fixations to the target, using videos of human gestures and human voice, and animated gestures and synthesized speech. Ambiguous, yet informative pointing gestures elicited attention and facilitated target identification, akin to verbal location descriptions. Moreover, target identification was superior when both pointing gestures and verbal location descriptions were used. These findings suggest that gesture not only operates as a context to verbal descriptions, or that verbal descriptions operate as a context to gesture, but that they complement one another in reference resolution.

Keywords: Gestures; Deixis; Deictic; Pointing; Multimodal communication; Referring expressions; Reference resolution; Eye tracking

Ever since McNeill's (1992) initial proposal that gesture and speech are part of a single integrated system, evidence has accumulated that they mutually interact in both production and, more recently, in comprehension (Kelly, Özyürek, & Maris, 2010). For example, everyday objects can be identified faster when their descriptions are accompanied by iconic gestures than when they are not (Riseborough, 1981), and iconic gestures affect verbatim recall for utterances (Kelly, Barr, Church, & Lynch, 1999). Gestures produced by speakers explaining a task affect motor actions of listeners subsequently doing the same task (Cook & Tanenhaus, 2009). Pointing gestures incongruent with linguistic information increase reaction times relative to congruent pairings (Langton, O'Malley, & Bruce, 1996) and affect pragmatic inferences about speaker intentions (Kelly et al., 1999).

Correspondence should be sent to Max M. Louwerse, Department of Psychology/Institute for Intelligent Systems, University of Memphis, 202 Psychology Building, Memphis, TN 38152. E-mail: mlouwerse@memphis.edu

Despite this increasing consensus, however, several blind spots remain. For instance, relatively little is known about how speech and gesture interact as referential communication unfolds (see Kelly, Manning, & Rodak, 2008). Studies investigating the time course of referring expressions have found that people can rapidly integrate linguistic and nonlinguistic information (Hanna & Brennan, 2007; Hanna & Tanenhaus, 2004; Richardson, Dale, & Kirkham, 2007; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002; Trueswell & Tanenhaus, 2005). Also, much research has focused on iconic gestures. But there is reason to believe that processes of gesture comprehension may vary according to the type of gesture involved. For one thing, gesture production varies according to gesture type. Iconic gestures decrease when partners are not mutually visible, whereas beat gestures do not (Alibali, Heath, & Myers, 2001). The fact that gesture production varies according to gesture type suggests that comprehension may as well. The different relation between different gesture types and speech may affect comprehension. Iconic gestures typically mimic particular aspects of their lexical affiliates, for example, physical properties or motion (Kelly et al., 2010), whereas deictic gestures focus attention of addressees on a subregion of shared visual space, thereby constraining interpretation by listeners (Bangerter, 2004). It may even be the case that different kinds of deictic gestures (e.g., pointing or gaze) affect comprehension differently, for instance depending on whether the gesture unambiguously identifies the target. Thus, although gaze may serve as a cue to the speaker's attention in a similar way as a pointing gesture (Hanna & Brennan, 2007), there is evidence that processing of gaze is more reflexive than processing of pointing cues, possibly because of the neural architecture specialized in eye processing (Friesen, Ristic, & Kingstone, 2004). Thus, there may be differences in the way pointing and gaze is processed in conjunction with verbal information.

There are currently no studies of the time course of integration of pointing gestures and speech in the resolution of referential expressions. Thus, the goal of the research reported here is to investigate how and when pointing gestures affect comprehension in real-time referential communication. How does referential communication unfold? It has been proposed that people typically identify a referent by describing its features so as to uniquely specify it among competitors within a domain (Olson, 1970). But with a large domain, such a strategy is not feasible, nor pragmatically appropriate. In real conversational situations, speakers do not design a perfectly unambiguous referential expression before initiating speech, and addressees do not wait for the speaker to do so. Rather, both speaker and addressee collaborate over several interactive turns to ground a contribution, often in a piecemeal, iterative fashion (Clark & Wilkes-Gibbs, 1986). An important initial component of this collaborative iterative design of referential expressions is circumscribing a subset of the domain. If speaker and addressee are mutually aware that they share a joint focus of attention, they will be able to identify the referent with less complex expressions. If the speaker knows the addressee is looking at a subset that contains only one red referent, *it's the red one* will be felicitous even though the whole domain may contain several red objects (Beun & Cremers, 1998). But speakers and addressees do not just try to guess where the other is looking (Hanna & Brennan, 2007). Rather, they try to actively manipulate each other's gaze, by using attention-focusing devices like pointing gestures (Clark, 2003) to substitute for

descriptions of spatial location (e.g., *the top left*), thereby reducing verbal effort in referring (Bangerter, 2004).

This study used eye-tracking methodology to test how and when ambiguous yet partly informative pointing gestures (e.g., pointing to the approximate region of the target) and verbal descriptions of spatial location (e.g., *on the right*) help addressees identify a target together with verbal descriptions identifying target features. Pointing gestures and location descriptions may facilitate identification by focusing addressee attention on a subdomain. If so, information enabling attention-focusing presented early in the time course should lead addressees to fixate on the target faster than when such information is not present. The key question is how fast such information can be processed. In naturalistic situations, gestures in production can be decomposed into a preparation, hold, and retract component (Kendon, 2004). We assume that it is possible to extract information from the pointing gesture starting with the onset of the hold, that is, when the gesture is immobile and focused on its target. Similarly, it should be possible to extract information from a location description soon after its utterance is complete. Therefore, we expected conditions where *one* of these two attention-focusing devices is present *in addition* to descriptions of target features to lead to earlier target identification relative to a baseline condition with *only* descriptions of target features. We also added a condition where *both* gestural information and location descriptions were present in addition to descriptions of target features.

In two experiments, we investigated the effects of location descriptions and pointing gestures on the time course of target identification during referring expressions. Participants viewed video clips on a computer monitor where a target was verbally identified by an unambiguous description of its features, for example, *John has a hat, a bowtie and glasses*. Their task was to identify the target. In the *baseline* condition, there was no other information. In the *pointing* condition, this feature description was accompanied by a pointing device (a human hand in Experiment 1, an arrow in Experiment 2) pointing ambiguously to the approximate target region. In the *location description* condition, this feature description was preceded by an ambiguous description of the approximate target location. Finally, in the *pointing and location description condition*, both attention-focusing devices were present.

1. Experiment 1

1.1. Method

1.1.1. Participants

Twenty-eight undergraduate students at the University of Memphis participated for course credit.

1.1.2. Materials

Participants saw 30 short movies where a human pointer described and/or pointed to a target among an array of faces (see Fig. 1). Only the pointer's arm was visible in the pointing

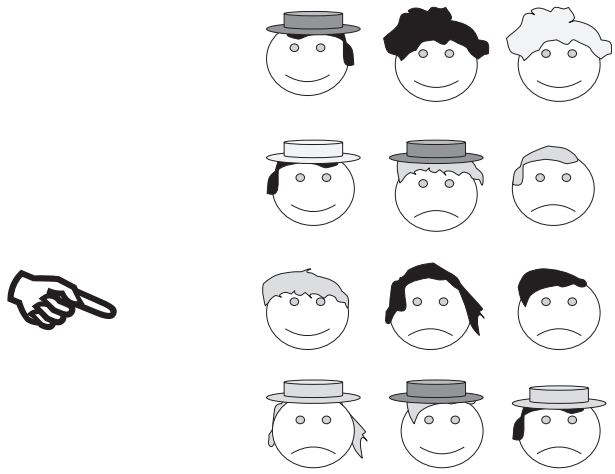

Fig. 1. Example of visual scene in Experiments 1 and 2 (in Experiment 1, the ClipArt hand is replaced by a human hand).

conditions. While describing the target, the pointer stood next to the array facing the participant and raised his arm in a natural motion with his fingertip being approximately at the same distance from the array of faces across conditions. Thus, the situation was two-dimensional. Arrays consisted of 12 smiley faces (three columns, four rows) differing in props (e.g., hat, bowtie, glasses) and emotion (happy, sad). Each face could be unambiguously described by three features. The factorial combination of the presence versus absence of a location description and the presence versus absence of pointing resulted in four conditions. In the baseline condition (no pointing, no location description), the feature description (specifying three features in sequence) was the only information available to the participants. In the other conditions, it was accompanied by the onset of a pointing gesture, or preceded by a location description, or both. Conditions and example instructions are shown in Table 1. Location descriptions specified a subset of the array (e.g., *in the middle*, *at the top*, *on the left*). Pointing gestures similarly highlighted a subset of the array (e.g., a row, see Fig. 1). Participants cycled through each condition five times in random order (each time with different targets and consequently different pointing gestures and/or descriptions), totaling 20 trials. In addition, they completed 20 filler trials. These fillers of short movies and unrelated text comprehension tasks (e.g., coherence relations tasks) were inserted as pilot stimuli for another study.

Table 1
Overview of experimental conditions in Experiments 1 and 2

	Pointing	No Pointing
Location description	[Pointing] + *John is in the middle with a happy face, dark hair and glasses*	*John is in the middle with a happy face, dark hair and glasses*
No location description	[Pointing] + *John has a happy face, dark hair and glasses*	*John has a happy face, dark hair and glasses*

1.1.3. Apparatus

Participants' eye movements were tracked using a Model 501 Applied Science Laboratories eyetracker (Bedford, MA; temporal resolution 60 Hz, spatial resolution <0.5°). Magnetic head tracking equipment was used to compensate for head movements. Participants were calibrated using a 9-point grid on a 1024 × 768 monitor both before and throughout the session to ensure reliable data. They were seated about 700 mm in front of the monitor.

1.1.4. Procedure

Participants watched each clip while their eye movements were recorded. After each clip, they selected the target by clicking one of the 12 faces.

1.2. Results and discussion

Because clips differed in duration (M = 8,050 ms, SD = 819 ms), we normalized data by calculating the percentage of fixations on the target relative to 40 time bins for each clip. Each bin lasted approximately 200 ms. Percentages were subjected to an empirical logit tranformation (Barr, 2008). We then conducted a 2 (presence/absence of location description) × 2 (presence/absence of pointing) mixed-effects model analysis on the empirical logit of fixation percentages to the target with experimental conditions and bin as fixed factors, and participants and items as random factors (Baayen, Davidson, & Bates, 2008). In addition to being more robust with regards to unequal cell sizes (Littell, Stroup, & Freund, 2002), the advantage of a mixed-effects model analysis over an analysis of variance (ANOVA) is that both differences between participants and differences between items are taken into account at the same time. The model was fitted using the restricted maximum likelihood (REML) estimation. F-test denominator degrees of freedom were estimated using the Kenward-Rogers adjustment for degrees of freedom, reducing the potential for a Type I error (Littell et al., 2002, p. 296).

As predicted, more fixations were made to the target for clips featuring pointing, $F(1, 26357.88) = 12.44$, $p < .001$. This was also the case for clips featuring location descriptions, $F(1, 26354.9) = 7.90$, $p = .004$. An interaction was found between pointing and location description, $F(1, 26349.79) = 6.82$, $p = .009$, showing that when a location description is missing, pointing leads to more fixations to the target than when a location description is present (Fig. 2B).[1] In Fig. 2A, earlier fixations to the target are clearly visible when an attention-focusing device (pointing, location description, or both) is present. The fixation curves are graphed in relation to the distributions of the onset and offset of the main parameters of the referential expressions: the pointing gesture, its hold, the speech stream, and the location description, as represented by box-and-whisker plots (boxes indicate the average onset and offset times and whiskers indicate standard errors).[2]

The fixation curves were fitted with five sigmoidal models (Gompertz relation, Logistic model, Morgan–Mercer–Florin [MMF] model, Richards model, and Weibull model). The fit of the MMF model (Seber & Wild, 2003)[3] was superior across the four data conditions (Table 2). The fitted curves can be seen in Fig. 2C. The derivative of the MMF function provides an estimate of the velocity of the eye fixations to the target (De'Sperati, 2003;

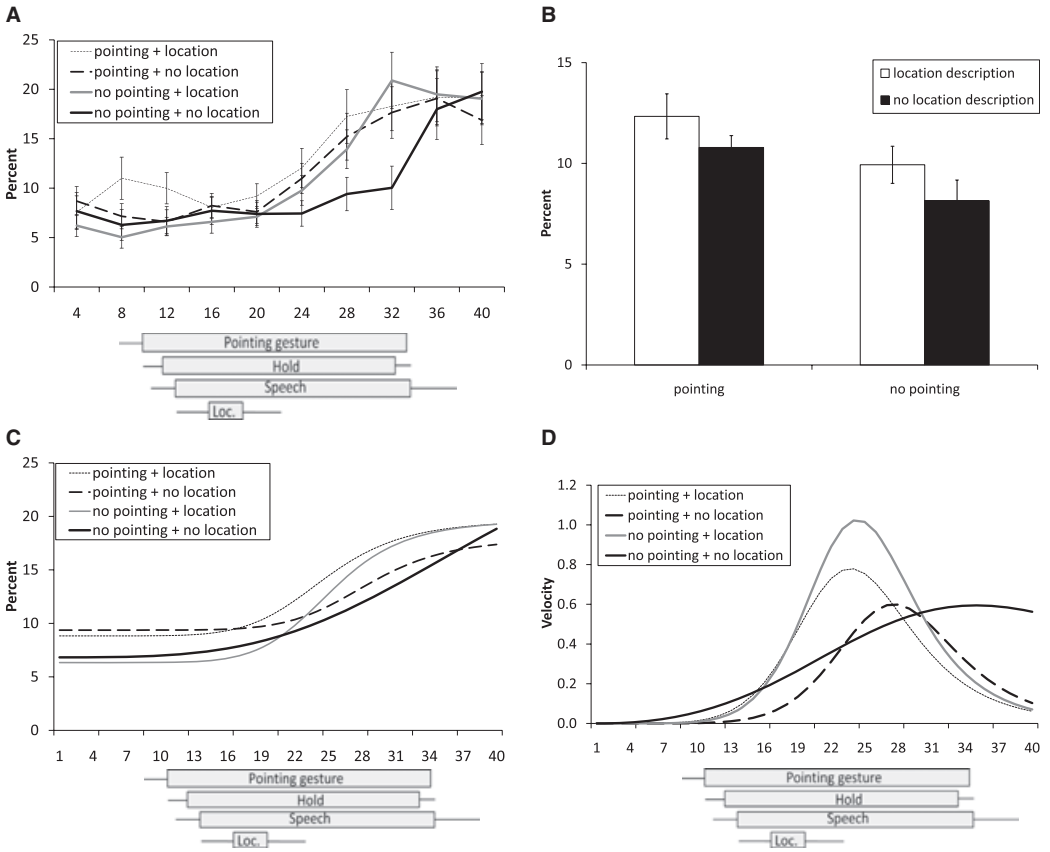

Fig. 2. Analyses of fixations to target, Experiment 1. (A) Percentage of fixations per time bin in each condition and distribution of onsets and offsets of key referring expression parameters. (B) Total percentage of fixations to target in each condition. (C) Morgan–Mercer–Florin (MMF) model–fitted curves of fixations per time bin in each condition. (D) Velocity of fixations to target per time bin in each condition.

Wojnowicz, Ferguson, Dale, & Spivey, 2009). The rate of change of position of the eye gaze is indicative of the moment of target identification, something that cannot be unambiguously determined by fixation time alone. The velocity curves are bell shaped and asymmetrical (Fig. 2D). In the *pointing* and *pointing and location description* conditions, the onset of the hold of pointing (bins 10–12) coincides with increase in velocity of fixations to the target. Also, in the *location* and *pointing and location description* conditions, the offset of the location description (bins 19–22) coincides with a slower rate of increase of velocity (i.e., a deceleration) of fixations toward target. In other words, the offset of the location description shortly precedes maximum velocity of fixations to the target. In the *baseline* condition, increase in velocity is more gradual and linear. This corresponds to the sequential integration of the three features in the feature description. In short, in the conditions where an attention-focusing device is present, fixations to the target take place very differently from the *baseline* condition. Interestingly, the peak velocity for fixations in the *pointing* condition

Table 2
Correlation coefficients, standard errors, and parameter coefficients for MMF model in Experiment 1

	Pointing + Location	Pointing + No Location	No Pointing + Location	No Pointing + No Location
r	.93	.94	.94	.87
SE	1.72	1.27	2.01	2.31
a	8.83	9.37	6.33	6.83
b	1.73E+10	2.81E+11	6.79E+10	6.26E+05
c	19.45	17.93	19.68	31.75
d	7.37	7.87	7.73	3.60

Note. MMF, Morgan–Mercer–Florin.

was attained later after this information became available (i.e., about 15 bins after the onset of the hold) than in the *location description* and *pointing and location description* conditions (i.e., about six bins after the offset of the location description). A mixed-effects analysis of the velocity data showed a significant interaction between the *pointing* condition and time bin, $F(1, 155) = 12.27$, $p = .001$, between the *location description* condition and time bin, $F(1, 155) = 30.03$, $p < .001$, but no three-way interaction, $F(1, 155) = .111$, $p = .74$.

Experiment 1 used videos with human-generated gestures and speech. These are naturally integrated because they share processing stages in production (De Ruiter, 2007). It is possible that subtle, uncontrolled differences in the stimuli like intonation, energy of the gesture, or speech-gesture synchrony may have affected eye gaze. We tried to eliminate these sources of variance in Experiment 2 by using artificially generated speech and gesture.

2. Experiment 2

Experiment 2 tested the same predictions as Experiment 1 except for two differences. First, an artificial environment was used where pointing gestures and linguistic expressions were generated independently from one another, using synthesized speech and a ClipArt hand. Second, a different eye tracker was used that stabilized the participant's head and used a higher sampling rate.

2.1. Method

2.1.1. Participants

Twenty-four undergraduate students at the University of Memphis participated for course credit.

2.1.2. Materials

As in Experiment 1, participants saw 30 short movies, with 12 smiley faces differing in props and emotion (see Fig. 1). A ClipArt hand with an extended index finger was used to point at the target; it appeared instantaneously at the onset of the referring expression and did not move. The voice was generated by the Rhetorical Systems' *r*Voice speech engine,

using the default intonation. The position of the faces and the hand, the feature description of the smiley faces using three distinctive features, and the location description (*left* and *right* vs. *top*, *middle*, and *bottom*) were the same as in Experiment 1, with each participant being exposed to all conditions.

2.1.3. Apparatus

An SMI iView X Hi-Speed eyetracker (Boston, MA) was used (temporal resolution 240 Hz, spatial resolution <0.5°). Participants were calibrated using a 9-point grid before and throughout the experimental session to ensure reliable data. The monitor was placed about 700 mm in front of the subject.

2.1.4. Procedure

The procedure was identical to that of Experiment 1.

2.2. Results and discussion

Even though speech synthesis and computerized pointing allowed for better control, video clips differed slightly in duration (M = 3,655 ms, SD = 342 ms). As in Experiment 1, we normalized the data by calculating the percentage of fixations on the target in 40 time bins for each clip, and computed the empirical logit for percentage fixations in each bin. Each bin lasted approximately 100 ms.

As in Experiment 1, we used a mixed-effects model fitted using REML and the Kenward–Rogers adjustment for degrees of freedom. The percentage of fixations on the target again was higher when pointing was present, $F(1, 18523.93) = 29.20$, $p < .001$, and when a location description was present $F(1, 18521.68) = 154.25$, $p < .001$ (Fig. 3B). As before, an interaction was found between pointing gestures and spatial descriptions, $F(1, 18508.16) = 7.20$, $p = .007$. These results replicate findings from Experiment 1. Earlier fixations to the target are clearly visible when an attention-focusing device is present (Fig. 3A).

As in Experiment 1, online effects of pointing gestures and location descriptions on target identification were investigated by fitting the data points with sigmoidal models. The MMF model best fitted the data (Table 3). The fitted curves can be seen in Fig. 3C. Following the procedure we used in Experiment 1, the derivative of the MMF function was computed to obtain an estimate of the velocity of the eye fixations to the target (Fig. 3D).

Because the pointer appeared instantaneously, gestural information was available from the onset of the referring expression. This information is used quickly: In the *pointing* and *pointing and location description* conditions, velocity of fixations to the target increases from the second bin onward. Also, in the *location description* condition the offset of the location description (bins 12–15) precedes maximum velocity of fixations to the target. As in Experiment 1, in the *baseline* condition, increase in velocity is more gradual and linear. This corresponds to the sequential integration of the three features in the feature description. That is, in the conditions where an attention-focusing device is present, fixations to the target take place very differently from the *baseline* condition. A mixed-effects analysis of the

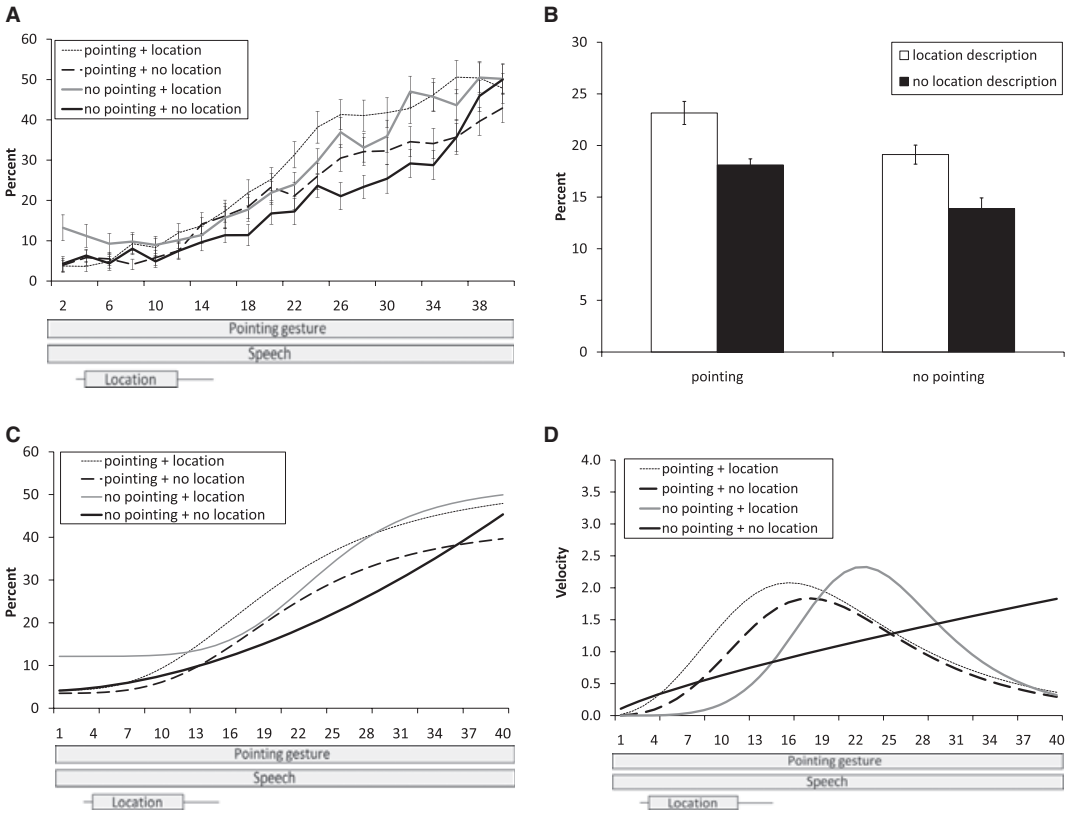

Fig. 3. Analyses of fixations to target, Experiment 2. (A) Percentage of fixations per time bin in each condition and distribution of onsets and offsets of key referring expression parameters. (B) Total percentage of fixations to target in each condition. (C) Morgan–Mercer–Florin (MMF) model-fitted curves of fixations per time bin in each condition. (D) Velocity of fixations to target per time bin in each condition.

velocity data showed a significant interaction between the *pointing* condition and time bin, $F(1, 155) = 5.12$, $p = .03$, the *location description* condition and time bin, $F(1, 155) = 202.98$, $p < .001$, and a marginally significant three-way interaction, $F(1, 155) = 3.89$, $p = .05$.

3. General discussion

This study investigated how and when ambiguous yet informative information about the approximate location of a target in a referring expression is used by participants. We found that both gestural (pointing) and verbal (location descriptions) cues affected the time course of fixations to targets. In conditions where this information is present, participants fixated to targets earlier than in the *baseline* (no pointing, no verbal descriptions) condition. The velocity of fixations is also very different in conditions where attention-focusing

Table 3
Correlation coefficients, standard errors, and parameter coefficients for MMF model in Experiment 2

	Pointing + Location	Pointing + No Location	No Pointing + Location	No Pointing + No Location
r	.99	.99	.99	.98
SE	2.13	1.68	2.76	2.72
a	4.17	3.50	12.14	4.08
b	9.58E+03	5.40E+04	3.15E+07	−2.61E+08
c	5.33E+01	4.33E+01	5.25E+01	−1.57E+07
d	3.05	3.58	5.42	1.77

Note. MMF, Morgan–Mercer–Florin.

information is present than in the *baseline* condition. Moreover, the peak in velocity of fixation to targets was temporally linked to the moment where this information became available (i.e., the onset of the hold for pointing gestures and the offset of the location description). In both experiments, it seems that the velocity of fixations peaked earlier after the offset of location descriptions than after the onset of the hold. It may in fact be easier to integrate location descriptions faster because they constitute intramodal information. Alternatively, using information in the pointing gesture may require fixations to the hand to relate the direction of the gesture to the target area. However, additional analyses revealed that the proportion of fixations on the gesture was quite low at around 7–9%. This corroborates previous research on gestures showing that addressees do not attend to gestures very much (Gullberg & Holmqvist, 2006). Even though the percentage of fixations was low, it was apparently sufficient for addressees to extract the relevant information—perhaps through the use of peripheral vision.

Our findings have two important implications. First, they address the question of how pointing gestures impact comprehension together with speech in shared task environments. The answer to this question is far from obvious (Goldin-Meadow, 1999), and it requires an experimental setup that moves away from paradigms employed to study how pointing is used to specify the referent of a deictic verbal expression. Such paradigms (e.g., Pechmann & Deutsch, 1982; Thompson & Massaro, 1994) typically study pointing used to disambiguate binary choices between stimuli. In this study, (a) the pointing information was ambiguous, and (b) linguistic information about target features was sufficient to identify the referent. We were therefore able to investigate how and when pointing helps addressees to focus attention on a referential subdomain (so that they do not need to systematically exclude all possible competitors linguistically). Although pointing is undoubtedly also used to specify a referent on its own, we argue that this is a special case of focusing attention, one where the subdomain is reduced to one referent. We have focused here on the more general case. The results add to an increasing body of evidence showing that integration of verbal and nonverbal information in complex visual worlds is fast and early (Trueswell & Tanenhaus, 2005).

Second, our findings have implications for the design of systems capable of generating multimodal references like embodied conversational agents (Cassell, Kopp, Tepper, Ferriman, & Striegnitz, 2007; Louwerse, Graesser, McNamara, & Lu, 2009; van der Sluis & Krahmer, 2007). In Experiment 1, human speech and gestures were used, possibly yielding

subtle effects in eye gaze due to intonation, energy of the gesture, and speech–gesture synchrony. Interestingly, the fixation time courses were similar in Experiment 2 in which a cartoon hand and synthesized speech were used. Thus, both natural and artificial referring expressions have similar effects on target identification.

Taken together, these findings show that gesture not only operates as a context to verbal descriptions, or that verbal descriptions operate as a context to gesture, but that they complement one another in reference resolution.

Notes

1. The percentage of fixations are proportions of fixations on the target. A fixation percentage of 20% might seem low, but is in fact high given that the target area is 11 times smaller than the nontarget areas.
2. All parameters were precisely timed by visual inspection using the linguistic annotation program ELAN. Onset and offset of individual words were determined by inspection of the wave form. Onset and offset of the gesture and each individual phase (preparation, hold, retract) were determined by frame-by-frame analysis of arm position. The onset of the preparation phase is determined by the frame where the pointer arm appears. The onset of the hold is determined by the frame where the arm stops moving. The offset of the hold is determined by the frame where the arm starts retracting. The offset of the retraction is determined by the frame where the arm disappears.
3. The function for the asymmetric sigmoidal four-parameter mathematical growth model (MMF) model is

$$y = \frac{ab + cx^d}{b + x^d}.$$

Acknowledgments

This research was supported by grants NSF-IIS-0416128 and SNSF-8210-061238. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding institutions. We would like to thank George Relyea for suggestions and comments on early drafts of this paper and Nick Benesh and Eric Mayor for assistance in coding the data. The usual exculpations apply.

References

Alibali, M. W., Heath, D. C., & Myers, H. 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.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, *15*, 415–419.

Barr, D. J. (2008). Analyzing ''visual world'' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*, 457–474.

Beun, R. J., & Cremers, A. H. M. (1998). Object reference in a shared domain of conversation. *Pragmatics and Cognition*, *6*, 111–142.

Cassell, J., Kopp, S., Tepper, P., Ferriman, K., & Striegnitz, K. (2007). Trading spaces: How humans and humanoids use speech and gesture to give directions. In T. Nishida (Ed.), *Conversational informatics* (pp. 133–160). New York: John Wiley & Sons.

Clark, H. H. (2003). Pointing and placing. In S. Kita (Ed.), *Pointing. Where language, culture, and cognition meet* (pp. 243–268). Hillsdale, NJ: Erlbaum.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*, 1–39.

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

De Ruiter, J. P. (2007). Postcards from the mind: The relationship between speech, imagistic gesture, and thought. *Gesture*, *7*, 21–38.

De'Sperati, C. (2003). Precise oculomotor correlates of visuo-spatial mental rotation and circular motion imagery. *Journal of Cognitive Neuroscience*, *15*, 1244–1259.

Friesen, C. K., Ristic, J., & Kingstone, A. (2004). Attentional effects of counterpredictive gaze and arrow cues. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 319–329.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Science*, *3*, 419–429.

Gullberg, M., & Holmqvist, K. (2006). What speakers do and what listeners look at. Visual attention to gestures in human interaction live and on video. *Pragmatics and Cognition*, *14*, 53–82.

Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, *57*, 596–615.

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science*, *28*, 105–115.

Kelly, S. D., Barr, D., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language*, *40*, 577–592.

Kelly, S. D., Manning, S., & Rodak, S. (2008). Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education. *Language and Linguistics Compass*, *2*, 1–20.

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

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge, England: Cambridge University Press.

Langton, S. R., O'Malley, C., & Bruce, V. (1996). Actions speak no louder than words: Symmetrical cross-modal interference effects in the processing of verbal and gestural information. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 1357–1375.

Littell, R. C., Stroup, W. W., & Freund, R. J. (2002). *SAS for linear models*. Cary, NC: SAS Publishing.

Louwerse, M. M., Graesser, A. C., McNamara, D. S., & Lu, S. (2009). Embodied conversational agents as conversational partners. *Applied Cognitive Psychology*, *23*, 1244–1255.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, *77*, 257–273.

Pechmann, T., & Deutsch, W. (1982). The development of verbal and nonverbal devices for reference. *Journal of Experimental Child Psychology*, *34*, 330–341.

Richardson, D. C., Dale, R., & Kirkham, N. Z. (2007). The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue. *Psychological Science*, *18*, 407–413.

Riseborough, M. G. (1981). Physiographic gestures as decoding facilitators: Three experiments exploring a neglected facet of communication. *Journal of Nonverbal Behavior*, *5*, 172–183.

Seber, G. A. F., & Wild, C. J. (2003). *Nonlinear regression*. Hoboken, NJ: Wiley Interscience.

van der Sluis, I., & Krahmer, E. (2007). Generating multimodal references. *Discourse Processes*, *44*, 145–174.

Spivey, M., Tanenhaus, M., Eberhard, K., & Sedivy, J. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, *45*, 447–481.

Thompson, L. A., & Massaro, D. W. (1994). Children's integration of speech and pointing gestures in comprehension. *Journal of Experimental Child Psychology*, *57*, 327–354.

Trueswell, J. C., & Tanenhaus, M. K. (Eds.) (2005). *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*. Cambridge, MA: MIT Press.

Wojnowicz, M. T., Ferguson, M. J., Dale, R., & Spivey, M. J. (2009). The self-organization of deliberate attitudes. *Psychological Science*, *20*, 1428–1435.