
Quotation as a Multimodal Construction

Natalia L. Blackwell\textsuperscript{a}, Marcus Perlman\textsuperscript{b}, and Jean E. Fox Tree\textsuperscript{a}

\textsuperscript{a}University of California, Santa Cruz. USA.

\textsuperscript{b}University of Wisconsin, Madison, USA.

Author Note

This research was supported by faculty research funds granted by the University of California Santa Cruz. We thank our many research assistants who aided in data collection and coding, with a special thanks to Annecy Majoros, Jordan Martin, Sarit Fassazadeh, Carly Johnson, and Aza Tetelman. Correspondence can be addressed to Natalia Blackwell, nblackwe@ucsc.edu, Marcus Perlman, mperlman@wisc.edu, or Jean E. Fox Tree, foxtree@ucsc.edu, Psychology Department, Social Sciences II, University of California, Santa Cruz, CA, 95064.
Abstract

Quotations are a means to report a broad range of events in addition to speech, and often involve both vocal and bodily demonstration. The present study examined the use of quotation to report a variety of multisensory events (i.e. containing salient visible and audible elements) as participants watched and then described a set of video clips including human speech and animal vocalizations. We examined the relationship between demonstrations conveyed through the vocal versus bodily modality, comparing them across four common quotation devices (be like, go, say, and zero quotatives), as well as across direct and non-direct quotations and retellings. We found that direct quotations involved high levels of both vocal and bodily demonstration, while non-direct quotations involved lower levels in both these channels. In addition, there was a strong positive correlation between vocal and bodily demonstration for direct quotation. This result supports a Multimodal Hypothesis where information from the two channels arises from one central concept.

Keywords: quotation, quotatives, demonstration, gesture, body, prosody
Quotation as a Multimodal Construction

Quotation is traditionally treated as a device used to report a person’s speech (Coulmas, 1986). Yet this characterization does not cover the wide range of ways that quotation is often used to demonstrate various kinds of events, often involving visible qualities as well as audible ones (Clark & Gerrig, 1990; Fox Tree & Tomlinson, 2008; Hudson, 1985; Waksler, 2001). The multimodality of quotation is illustrated in the following examples of (1) a vocal demonstration of an audible event, (2) a bodily demonstration of a visible event, and (3) a combined vocal and bodily demonstration of a multisensory event:

(1) The car engine went [brmbrm], and we were off (Clark & Gerrig, 1990; from Hudson, 1985).

(2) I got out of the car, and I just [demonstration of turning around and bumping his head on an invisible telephone pole] (Clark & Gerrig, 1990; from Hudson, 1985).

(3) He was like opening his drawers but they were like stuck so he was like, “Raah aah” [growling sounds, while pretending to pull on a drawer] (Fox Tree & Tomlinson, 2008: 100).

Here we examine how vocal and bodily demonstrations are coordinated in direct and non-direct quotations, including quotations that followed three common quotative devices, be like, go, and say. Our investigation sheds light on quotation as a multimodal construction, and more generally, on the nature of language as a multimodal system of communication.
**Quotation.** The canonical, *direct* form of quotation allows a speaker to convey another person’s words such as in the following example (this and all further examples in this paper are from the current corpus):

(4) He said, “After a while, crocodile.”

Quotation can also be used for other, non-speech sounds, such as the pre-linguistic babbling of an infant, the vocalization of a nonhuman animal, or the crack of an inanimate firecracker:

(5) The babies were going, “Ga ga ga da da.”
(6) The cows were like, “Mooo.”
(7) It goes, “Pop pop pop pop pop pop pop pop.”

Although direct quotations are sometimes thought of as veridical reports of what was said, they are not necessarily verbatim (Blackwell & Fox Tree, 2012; Fox Tree & Tomlinson, 2008; Wade & Clark, 1993; and for discussion, see Clark & Gerrig, 1990; Romaine & Lange, 1991). Instead, they appear to be used to create imagined experiences (Pascual, 2014), or to recreate for the listener “what it would be like to hear, see, or feel what the original speaker did” (Wade & Clark, 1993: 818). Quotations are often associated with dramatic presentation and elevated levels of sensorimotor imagery. For example, speakers often produce direct quotations for dramatic effect at peak points in a narrative (Li, 1986), and in turn, listeners find narratives to be more dramatic and memorable when presented as direct, first-person speech (Schiffrin, 1981; Tannen, 1989). In addition, reading direct quotations, but not indirect quotations, activated areas of the brain related to voices (Yao, Belin, & Scheepers, 2011). Actual reading behavior was also
affected – reading rate was faster when quoting a fast-paced speaker as compared to a slow-paced speaker (Yao & Scheepers, 2011).

In English and other languages, direct quotations are frequently introduced by particular lexical devices, or *quotatives*, such as *tell*, *ask*, *say*, *be like*, and *go*. A number of factors contribute to the choice of a particular device, including the nature of the quoted content as well as social factors. Compared to *say*, the use of *be like* is associated with looser reports of the original quoted content and with the expression of attitude or emotion (Andersen, 1998, 2000; Blyth, Recktenwald, & Wang, 1990; Romaine & Lange, 1991), although both of these proposals have been challenged (Blackwell & Fox Tree, 2012; Fox Tree & Tomlinson, 2008). Compared to *be like*, the use of *say* is associated with higher status speakers being quoted and higher status addressees (Blackwell & Fox Tree, 2012). Both *be like* and *go* are associated with the reporting of non-speech sounds and also demonstrable actions and gestures (Butters, 1980; Buchstaller, 2001; Fox Tree & Tomlinson, 2008). In some cases, quotations are produced without any introductory device to frame the quoted material (i.e., *zero quotatives*). The lack of any verb requires speakers to use more vocal modulation to signify the act of quotation, and zero quotatives are often used when conveying a more dramatic representation of the quoted speech (Mathis & Yule, 1994).

**Gesture and Quotation.** Previous researchers have shown a hand-in-hand relationship between the amount of information conveyed through the words of an utterance and the amount of information conveyed through co-speech gesture (de Ruiter, Bangerter, & Dings, 2012; So, Kita, & Goldin-Meadow, 2009). Utterances that are more informative by words also tend to be
more informative by gesture. However, under certain circumstances there can also be a trade-off between speech and gesture. When speech is more effortful, such as when referring a partner to mutually visible targets that are more difficult to distinguish verbally, gesturing takes over some of the cognitive load. When gesturing is more effortful, as when the targets are farther away, speech takes over (Bangerter, 2004; de Ruiter, 2006; van der Sluis & Krahmer, 2007).

The production of quotations in particular is associated with a high rate of gesturing, and it is often coordinated with distinctive patterns of eye gaze and body orientation (Goodwin, 2007). The production of quotations is also associated with vocal demonstrations, so that words are quoted with additional vocal qualities, such as enacted lisps or drawls, and non-speech sounds are quoted as vocal imitations of the originals (Clark & Gerrig, 1990; Wade & Clark, 1993). Yet, when speakers use quotations to demonstrate an event, especially non-speech events, the question remains as to how they coordinate the vocal and bodily channels. That is, we know that multimodal communication can take place in a variety of ways, such as hand gestures and words indicating different aspects of the meaning of a concept (Kendon, 2014). We ask how vocal demonstration and bodily demonstration coordinate with each other.

A Multimodal Hypothesis predicts that vocal and bodily demonstrations arise from a single underlying representation of the quoted content, one that does not distinguish between sensory modality. Thus demonstration through the vocal and bodily channels would both rise concomitantly with a general rise in the overall degree of demonstration. Alternatively, according to a Trade-Off Hypothesis, a demonstration is produced preferentially through one channel or the other. Quotation potentially poses an increased cognitive load on the speaker who
must reproduce a relatively detailed account of the quoted event, often in the context of presenting a story. Thus the cognitive effort required to produce a vocal demonstration may detract from the level of bodily demonstration, and vice versa, such that the more one type of demonstration is used, the less the other is used. Finally, according to an Independent Hypothesis, there might be no relationship between these two modes of demonstration.

**Current Study.** In the current study, we examined the use of quotation as a multimodal construction to report an expanded variety of speech and speech-like events, including accented human speech, pre-linguistic babbling of babies, speech-like vocalizations produced by animals, and speech produced by robots. Each of the events also contained salient visible elements. The main purpose was to investigate how speakers coordinate demonstration in the vocal and bodily modalities. This would determine which of the three hypotheses, Multimodal, Trade-Off, or Independent, best describes the way in which such kinds of events are quoted. In addition, we investigated whether there was a relationship between the level of demonstration and the selection of particular quotation devices.

**Method**

**Participants**

One hundred and thirty-seven undergraduates at the University of California, Santa Cruz, participated in the experiment in exchange for course credit.

**Materials**

Participants watched eighteen short YouTube videos (40 secs - 2 min 20 secs in length) involving an assortment of quotable sounds and actions. The videos contained a variety of
vocalizations by humans (adult and infant) and nonhuman animals, sounds produced by inanimate objects, and a variety of movements.

**Procedure**

Participants described each video to a confederate addressee directly after viewing it. They were instructed to be descriptive and entertaining. The confederate was allowed to laugh and provide appropriate back channels. The descriptions were video-recorded and transcribed for analysis. Individual descriptions were between 20 seconds and 4 minutes ($M = 1.73 \text{ min}, SD = 3.21 \text{ min}$).

**Analysis**

Analysis of the video-recordings proceeded in two steps. In the first step, stretches of recordings that met certain criteria were selected. In the second step, ratings of these stretches of recording (*clips*) were obtained.

**Quotation Selection.** To eliminate variability from the quotation of different content within a video, we focused analysis on descriptions of a single quotable event for each clip (the primary audible event) and included only quotations related to that particular event. To eliminate video-quotative confounds, we further focused our analyses on the clips that elicited uses of each of three verbal quotatives: *be like*, *go*, and *say* (*just* was scarcely used and not evaluated further). If *goes* was always used for one video and *says* for another, we would not be able to tease apart the contribution to demonstration that resulted from the use of a particular quotative versus a particular video. Ten clips did not elicit instances of each of these quotatives, leaving eight clips for analysis: (1) a dog howling what sounded like the words *I love you*, (2) a parrot producing
imitations of human speech and a tiger’s roar, (3) another parrot counting to ten, (4) two infants who appeared to be talking to each other, (5) a talking robot, (6) a speaker pretending to have a French accent, (7) a woman falling during a news interview about grape stomping, and (8) a girl asking her father for ketchup. We included all descriptions of the targeted events from the eight videos, even when they did not contain syntactically-identifiable quotation. Thus our corpus included canonical direct quotation constructions with quotatives and zero quotatives, and also non-direct reporting of the events. This latter category included indirect quotation (infrequent with non-speech) and descriptions, which could both be accompanied by bodily demonstration (see Examples 8-10; note that in example 8 the like is a discourse marker like, not a quotative like, Fox Tree, 2006).

(8) The kittens were like meowing to each other.

(9) They’re talking about how they’re going to make wine.

(10) It was just counting up from 1 to 10.

Thus the reporting or describing of any of the 8 events listed, with or without quotation devices, was included in the analysis.

Ratings. From the resulting corpus, for each clip we had about 35 direct quotations (using be like, say, and go, and zero quotatives) and 10 non-direct descriptions, depending on how many reporters included the critical event in their retellings. The quotations and descriptions were rated for the degree of vocal and bodily demonstration. The ratings ranged on a scale of 1 (not very demonstrative) to 5 (very demonstrative) and were made by research assistants who were naïve to the hypotheses of the study. To facilitate independent ratings across the two
modalities, ratings for vocal demonstration were performed on the audio only, while bodily demonstration ratings were performed on the video without audio. Quotatives were excised in order to prevent bias.

As training, raters were first shown several example clips of participants’ descriptions that illustrated a range of levels of vocal and bodily demonstration. Raters were instructed to first listen to or watch the full set of clips for a particular event in order to determine the range of demonstrations possible, and then to judge each retelling against the set. It was explained that vocal demonstrations included any way the speaker modulated his or her voice, deviating from the rest of the speech, in order to show something about the event. Bodily demonstration included any movements that could be thought of as using the body to show the event. Raters were instructed to take the whole body into consideration and to attend to both the size of movements and the number of movements produced. In order to not bias the raters, detailed instructions defining bodily demonstration were not provided, thus ratings could include postural movements, gaze movements, or facial expressions in addition to hand motions. Raters practiced until they reached a high level of inter-rater reliability with clips that were not used in the analyses. The actual clips were each rated by two raters (a total of six raters), with good inter-rater agreement for both vocal ($\kappa_w = .79, 95\% \ CI, .75 \ to \ .83$), $p < .001$) and bodily ($\kappa_w = .78, 95\% \ CI, .74 \ to \ .82, p < .001$) demonstration.

Because not all participants produced all five quotative devices, a mixed model linear regression was performed, with the device as a fixed factor, to determine the effect of the device type on the level of vocal and bodily demonstration.
**Results**

Table 1 gives examples of the vocal and bodily demonstrations produced by participants after viewing each of the original videos. For instance, when describing a video of a young girl asking her father for ketchup, participants often demonstrated her high-pitched voice to quote her request, “Daddy pass the ketchup,” often as they gazed upward as if towards the girl’s imagined father. Or when describing a dog howling what sounds like “I love you,” participants howled the words, and like the dog, raised their head upwards in a howling motion.

<table>
<thead>
<tr>
<th>Original Video</th>
<th>Quoted speech/sound</th>
<th>Bodily demonstration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Talking Dog</td>
<td>“I love you”</td>
<td>Raises head in a howling motion</td>
</tr>
<tr>
<td>Talking Parrot</td>
<td>“I’m a double yellow-headed Amazon” “Grrr”</td>
<td>Bobs head</td>
</tr>
<tr>
<td>French Accent</td>
<td>“Don’t you know we French men are loyal?”</td>
<td>Clenches hand (like the speaker)</td>
</tr>
<tr>
<td>Grape Stomping</td>
<td>“Ow, ow, ow”</td>
<td>Writhes in pain[4]</td>
</tr>
<tr>
<td>Ketchup Asking</td>
<td>“Daddy pass the ketchup”</td>
<td>Gazes upward as if to daddy</td>
</tr>
<tr>
<td>Babies Babbling</td>
<td>“Da da da da”</td>
<td>Shakes arms (like the babies)</td>
</tr>
<tr>
<td>Counting Parrot</td>
<td>“One, two, six”</td>
<td>Bobs head</td>
</tr>
<tr>
<td>Talking Robot</td>
<td>“Let’s go to the rocket ship”</td>
<td>Bends arms in a rigid position</td>
</tr>
</tbody>
</table>

Table 2 shows the proportion of non-direct quotations found in the resulting corpus (from which the proportion of direct quotations can be derived by adding to 100%; for example, 19.2% of the Dog clip were indirect and 80.8% were direct). Table 2 also shows, within the direct quotations, the distribution of quotatives used (such that all direct quotations taken together add to 100%).

Across all clips, the quotation device *be like* was used most frequently overall, 42.5%, followed by *say*, 22.6%, and *go*, 18.1%. Zero quotatives accounted for 6.2% of quotatives and *be just* for 3.9%. Other devices were infrequent, apart from some more specific reporting verbs associated
with particular videos: *ask*, elicited by the ketchup video (e.g. “the little girl asks, ‘can I have some ketchup?’”) and *count*, elicited by the counting parrot (the parrots counts, “one, two...”).

<table>
<thead>
<tr>
<th>Clip</th>
<th>Type of sound</th>
<th>Direct quotations</th>
<th>Non-Direct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td><em>Be like</em></td>
<td><em>Say</em></td>
</tr>
<tr>
<td>Dog Talking Parrot</td>
<td>Non-Human Speech</td>
<td>42.0</td>
<td>19.4</td>
</tr>
<tr>
<td>French</td>
<td>Human Speech</td>
<td>40.3</td>
<td>32.8</td>
</tr>
<tr>
<td>Stomping</td>
<td>Human Speech</td>
<td>68.1</td>
<td>10.6</td>
</tr>
<tr>
<td>Ketchup</td>
<td>Human Speech</td>
<td>42.9</td>
<td>11.4</td>
</tr>
<tr>
<td>Babies Counting Parrot</td>
<td>Human Non-Speech</td>
<td>34.1</td>
<td>10.7</td>
</tr>
<tr>
<td>Robot</td>
<td>Non-Human Speech</td>
<td>64</td>
<td>12</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>42.5</td>
<td>22.6</td>
</tr>
</tbody>
</table>

The distribution of quotatives differed from the distribution observed in a corpus of autobiographical narratives and conversations, where 92% of quotation devices were *like*, 7% were *said*, and 2% were *goes* (Fox Tree & Tomlinson, 2008; see also Blackwell & Fox Tree, 2012, who found a similar distribution with autobiographical narratives and narrative retellings). The clip-retelling technique increased the chances that a non-*like* quotative was used in direct quotations. Part of the reason for this could be because the clip-retelling technique encouraged quoting of a wider range of behaviors than narratives and conversations. For example, if sounds are more likely to be introduced by *goes* (as suggested by Romaine & Lange, 1991), and sounds are more likely to be quoted in clip retellings (current study) than in autobiographical narratives,
narrative retellings, and conversations (cf. Blackwell & Fox Tree, 2012; Fox Tree & Tomlinson, 2008), then there would be more quotations with *goes* in the current study.

Of the total 137 participants, 89 produced at least two different types of quotation devices and one non-direct quotation in the eight targeted videos. The mean ratings for vocal and bodily demonstration are given in Table 3.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Mean Ratings (SD) of Vocal and Bodily Demonstrations on a Scale from 1(not at all) to 5(very)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
</tr>
<tr>
<td>Zero</td>
<td>23</td>
</tr>
<tr>
<td>Be Like</td>
<td>77</td>
</tr>
<tr>
<td>Go</td>
<td>68</td>
</tr>
<tr>
<td>Say</td>
<td>81</td>
</tr>
<tr>
<td>Non-Direct</td>
<td>89</td>
</tr>
</tbody>
</table>

A mixed model analysis, with device type (*be like*, *go*, *say*, zero, and non-direct) and demonstration type (vocal and bodily) as fixed factors, revealed a significant effect of device type on the level of demonstration, $F(4, 38.9) = 26.3, p < .001$. No difference was found between the demonstration types, $F(1, 635.1) = 2.15, p = .07$, and no interaction was found between device type and demonstration type, $F(4, 635.1) = 0.52, p = .47$. For both vocal and bodily demonstration, the three highest ratings were for zero, *be like*, and *go*, then *say*, then non-direct. For vocal demonstrations, post-hoc pairwise comparisons with Bonferroni adjusted levels of .005
(.05/10) showed similar ratings for zero, be like and go (no difference between zero and be like, $t(19) = .30, p = .77$, or between be like and go $t(56) = .85, p = .40$). There was, however, a reliable difference between go and say, $t(62) = 6.34, p < .001$, and between say and non-direct, $t(80) = 4.01, p < .001$. Similarly, levels of bodily demonstration were also higher for be like, go, and zero quotatives compared to say and non-direct quotations. Here, the level of demonstration was higher for zero than be like, $t(19) = 2.67, p = .02$, and higher than go, $t(19) = 2.83, p = .01$. There was no difference between be like and go, $t(56) = .16, p = .87$, but there was a difference between be like and say, $t(69) = 3.53, p = .001$, and between say and non-direct, $t(80) = 3.49, p = .001$.

The relationship between the two types of demonstrations were then assessed for direct and non-direct quotation. In direct quotations, vocal and bodily demonstrations were strongly correlated, $r(249) = .59, p < .001$. Correlations for each video, with Bonferroni adjusted levels of .006 (.05/8), also revealed a positive association between these two types of demonstration for every clip (see Table 4).

<table>
<thead>
<tr>
<th>Clip</th>
<th>Test Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Talking Dog</td>
<td>$r(38) = .60, p &lt; .001$</td>
</tr>
<tr>
<td>Talking Parrot</td>
<td>$r(32) = .68, p &lt; .001$</td>
</tr>
<tr>
<td>French Accent</td>
<td>$r(34) = .73, p &lt; .001$</td>
</tr>
</tbody>
</table>
For each device, with Bonferroni adjusted levels of .01 (.05/5), vocal and bodily demonstrations were also strongly correlated. They were not correlated for non-direct demonstrations (see Table 5).

Table 5

Correlations Between Vocal and Bodily Demonstrations
for Each Quotative Device and for Non-Direct Quotation

<table>
<thead>
<tr>
<th>Type of Quotation</th>
<th>Test Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero</td>
<td>$r(23) = .46, p = .003$</td>
</tr>
<tr>
<td>Be like</td>
<td>$r(77) = .54, p &lt; .001$</td>
</tr>
<tr>
<td>Go</td>
<td>$r(68) = .55, p &lt; .001$</td>
</tr>
<tr>
<td>Say</td>
<td>$r(81) = .44, p &lt; .001$</td>
</tr>
<tr>
<td>Non-direct</td>
<td>$r(89) = .14, p = .29$</td>
</tr>
</tbody>
</table>

Discussion
Quotation appears to be a multimodal construction. When speakers use direct quotations, they generally produce a high level of demonstration in both the vocal and bodily channels. Moreover, the level of demonstration in each channel is correlated. When speakers use more vocal demonstration, they also use more bodily demonstration when producing direct quotations. The highest amount of demonstration was found after *be like, go,* and zero quotatives. Of the direct quotation devices, *say* was associated with less vocal and bodily demonstration than the other three. Non-direct descriptions were the least demonstrative in the vocal channel, and, notably, were also the least demonstrative in the bodily channel.

The overall distribution of the quotatives obtained with the video-reporting task used here was different from the distribution found in narratives and conversations (cf. Blackwell & Fox Tree, 2012; Fox Tree & Tomlinson, 2008). The prevalence of *be like* was similar to the earlier observations. But at 18%, the use of *go* was considerably higher than the low percentage (2%) found earlier. One conclusion is that rather than falling out of use, as suggested by *like’s* becoming more popular and *goes’* becoming less popular over a twenty-year period (Fox Tree & Tomlinson, 2008), *go* may have become specialized for highly demonstrative quotations. Other evidence provided here suggests that *go* is especially useful for vocal demonstrations.

Across each of the four quotatives (*be like, go, say,* and zero quotatives), as well as across each of the videos, vocal and bodily demonstrations showed a strong positive correlation. While non-direct descriptions did not show this correlation, the overall level of demonstration was generally quite low for both channels. Previous studies have suggested a trade-off between gesture and speech under challenging conditions of speaking or gesturing, in which the speaker
favors the easier mode of communication. In the present study, we considered the relationship of the vocal channel with gesture in the context of vocal demonstration, rather than speech. For this case, our findings support the Multimodal Hypothesis: a high correlation between activity in the vocal and bodily channels. When performing the demonstrative part of a quotation, the vocal and bodily channels are naturally integrated together into a unified, multimodal form. Indeed, previous work has shown that even in situations focused on sound, such as in the use of quotation in music instruction where physical motion has no literal meaning, gestures accompanying non-lexical vocalizations are still produced (Tolins, 2013). The lack of a correlation in non-direct speech could be a reflection of one manifestation of the Multimodal Hypothesis, with low vocal and low bodily demonstration.

We conclude with some broader consideration of the significance of quotation for investigating the relationship between language and gesture. A fundamental point of contention among cognitive scientists concerns whether language and gesture are complementary parts of a single system or separate systems of communication (e.g., McNeill, 1992; Levinson & Holler, 2014). Quotation offers a unique perspective into this relationship. In the formation of a quotation construction, the lexical selection of a quotative appears to be connected to the nature of the demonstration the speaker intends to perform. We found that some quotatives (e.g. *go* and *be like*) are associated with an increased overall level of demonstration in the following quoted content, whereas others (e.g. *say*) are associated with lower levels of demonstration. In this way, quotation bridges language and gesture, and connects the linguistic and imagistic processes that are involved in spoken communication. Moreover, the present study shows that
quotation, particularly as it uses lexical cues like *be like, go*, and zero quotatives, is aptly considered multimodal, integrating the vocal and bodily channels into a single demonstration. Researchers have given considerable attention to how signed languages have the means to describe an event through a grammatical integration of conventional morphemic forms and analogical demonstrations (e.g., classifier constructions; Emmorey, 2003; Liddell, 2003). Likewise, spoken languages may be seen to accomplish the integration of linguistic and analogical components through the use of the multimodal construction of quotation.
References


