Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 72 (2015) 160–175

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

On what happens in gesture when communication is unsuccessful

Marieke Hoetjes [*], Emiel Krahmer, Marc Swerts

Tilburg Center for Cognition and Communication (TiCC), Tilburg University, The Netherlands

Received 9 July 2014; received in revised form 16 April 2015; accepted 2 June 2015
Available online 6 June 2015

Abstract

Previous studies found that repeated references in successful communication are often reduced, not only at the acoustic level, but also in terms of words and manual co-speech gestures. In the present study, we investigated whether repeated references are still reduced in a situation when reduction would not be beneficial for the communicative situation, namely after the speaker receives negative feedback from the addressee. In a director–matcher task (experiment I), we studied gesture rate, as well as the general form of the gestures produced in initial and repeated references. In a separate experiment (experiment II) we studied whether there might (also) be more gradual differences in gesture form between gestures in initial and repeated references, by asking human judges which of two gestures (one from an initial and one from a repeated reference following negative feedback) they considered more precise. In both experiments, mutual visibility was added as a between subjects factor. Results showed that after negative feedback, gesture rate increased in a marginally significant way. With regard to gesture form, we found little evidence for changes in gesture form after negative feedback, except for a marginally significant increase of the number of repeated strokes within a gesture. Lack of mutual visibility only had a significant reducing effect on gesture size, and did not interact with repetition in any way. However, we did find gradual differences in gesture form, with gestures produced after negative feedback being judged as marginally more precise than initial gestures. The results from the present study suggest that in the production of unsuccessful repeated references, a process different from the reduction process as found in previous studies in repeated references takes place, with speakers appearing to put more effort into their gestures after negative feedback, as suggested by the data trending towards an increased gesture rate and towards gestures being judged as more precise after feedback.

Keywords: Gesture; Speech; Repeated references; Negative feedback

1. Introduction

People often refer to objects and persons during a communicative exchange. In many cases, the same target is referred to repeatedly in the discourse, and these references may be multimodal, using both speech and manual co-speech gesture. It is well established that repeated references in successful communication tend to be reduced variants of initial references, consisting of less words and less gestures. For example, a speaker who wants to point out a particular person for an addressee might produce an initial description such as “that tall girl with the long blond hair”, accompanied by two gestures, first one indicating the height of the girl, followed by another one indicating the length of the girl’s hair. Later on in the conversation, the speaker might refer back to the same girl by saying “the tall girl from before”, accompanied by only one gesture, say, indicating the girl’s height.

These reduction effects have been explained in terms of increased common ground (e.g., Clark and Wilkes-Gibbs, 1986; Galati and Brennan, 2014; Gerwing and Bavelas, 2004; Holler and Stevens, 2007; Jacobs and Garnham,

[*] Corresponding author at: Room D 404, PO Box 90153, 5000 LE Tilburg, The Netherlands. Tel.: +31 13 466 2918.

E-mail addresses: m.w.hoetjes@tilburguniversity.edu (M. Hoetjes), e.j.krahmer@tilburguniversity.edu (E. Krahmer), m.g.j.swerts@tilburguniversity.edu (M. Swerts).

http://dx.doi.org/10.1016/j.specom.2015.06.004
0167-6393/© 2015 Elsevier B.V. All rights reserved.

2007). The initial description introduces an entity in common ground, after which a reduced reference can be sufficient. The emergence of common ground is the result of a process often referred to as information grounding (Clark and Schaeffer, 1989; Traum, 1994), and generally understood as involving two phases: a presentation phase, in which a speaker sends a message to the addressee, and an acceptance phase, in which the addressee signals whether the message came across in good order or not. If our addressee knows which tall, long-haired girl the speaker is referring to, he[1] can signal this using a positive "go on" signal (using the terminology of Krahmer et al., 2002). This can, for example, be an explicit backchannel cue such as "OK", but it may also be a more implicit signal, because the addressee correctly identifies the target girl, e.g., by looking at her.

Now, consider what would happen if the initial reference is somehow not successful, which our addressee would indicate during the acceptance phase using a negative, "go back" signal (e.g., "Sorry, which girl?"). Then, how would our speaker realise her second, repeated reference to said girl? We know from other studies that speakers tend not to reduce their utterances (in terms of number of words or articulatory effort) in response to negative feedback, but we know remarkably little about whether, and if so, how, speakers' gestures would change. To the best of our knowledge only a handful of earlier studies asked this question, of which Holler and Wilkin (2011) is arguably the most detailed. However, these authors present their work as "a first glimpse of speakers' gestural behaviour in response to addressee feedback" (Holler and Wilkin, 2011, p. 3534), and point out that more work is "urgently needed" (ibid.).

In the present study we address the above questions by comparing gestures produced in initial references with those in repeated references following negative feedback. The experiments that were conducted for this purpose are based on the experimental paradigm of our previous work on successful repeated references (Hoetjes et al., 2011, 2015). As in this previous work (as well as in various other studies, including the aforementioned Holler and Wilkin, 2011), we concentrate on two aspects: the gesture rate and the qualitative form of the gestures. Before describing our current study in detail, we provide an overview of relevant background literature.

2. Background

2.1. Reduction in successful repeated references

Repeated references occur in discourse whenever a particular person or object is mentioned or described more than once. These references are never exactly the same. The differences in the ways in which references are realised are not only due to naturally occurring variability in speech, but are also influenced by the mere fact that the information status of the referent changes when it gets repeated. For instance, when an object is mentioned a second time, it already belongs to the discourse model of speaker and addressee, and can be assumed to be common ground (that is, when communication was successful). Research has found that when information is given or predictable, such as is the case in repeated references and increased common ground, speech is often reduced.

For example, Lieberman (1963) found that words produced in contexts in which they were predictable, had a shorter duration, and a lower pitch peak (F0). In addition, they were less intelligible when they were taken out of context. In a similar vein, references to given information have been found to be less intelligible when taken out of context and presented in isolation (e.g., Bard et al., 2000; Fowler and Housum, 1987), and to have a shorter duration and a lower pitch peak (e.g., Aylett and Turk, 2004; Brown, 1983; Fowler and Housum, 1987; Lam and Watson, 2010), than references to information that is new in the discourse.

Reduction in repeated references at the lexical level has also been well established. For example, Clark and Wilkes-Gibbs (1986) showed that when speakers repeatedly (and successfully) refer to the same object, they lexically reduce their references (e.g. from an initial description such as "a person who's ice skating, except they're sticking two arms out in front", to a sixth description of the same figure as "the ice skater", Clark and Wilkes-Gibbs, 1986, p. 12). This robust finding has often been explained in terms of the creation of a conceptual pact (Brennan and Clark, 1996), which occurs as more common ground emerges between speakers.

These findings relate to spoken language, but human speakers are known to produce speech in tandem with a variety of visual cues, of which manual gestures are our main focus of attention in this study. Such manual speech-accompanying or co-speech gestures (which we will call gestures for short) can generally be defined as symbolic movements of the arms and hands that people produce when they speak (Kendon, 1980, 2004; McNeill, 1992). Most researchers agree that there is a close, co-expressive relationship between speech and gesture (Kendon, 1972, 1980, 2000, 2004; McNeill, 1985, 1992; McNeill and Duncan, 2000), with speech and gesture arguably going "hand-in-hand" (e.g., Kita and Özyürek, 2003; So et al., 2009). To take one, more or less arbitrary, example, consider the study reported by So et al. (2009), who asked English speakers to retell stories to an experimenter. So and colleagues found that speakers often used gestures to identify a referent in the story, by producing it in the same location used for the previous gesture for this referent. However, importantly, they did this most often when the referent was also uniquely specified in the accompanying speech. This led these authors to conclude that for

[1] Throughout this paper, 'she' will be used to indicate the speaker, and 'he' to indicate the addressee.

referential identification, speech and gesture indeed appear to go hand-in-hand.

Based on this, one could hypothesise that reduction in speech during successful communication is accompanied by reduction in gesture. This is indeed what a number of studies have investigated, and to some degree the results are consistent with this hypothesis. For instance, it is generally found that repeated multimodal references contain fewer gestures than initial ones (e.g., de Ruiter et al., 2012; Holler et al., 2011; Levy and McNeill, 1992; Masson-Carro et al., 2014), just as they contain fewer words. However, when looking at the ratio of gestures to words a more complex picture emerges. Gesture rate (often computed as the ratio of gestures per 100 words, although various alternatives have been proposed, see Hoetjes et al., 2015, for discussion) has a long tradition in gesture research, going back to, at least, Cohen and Harrison (1973). It has frequently been used as a dependent variable in gesture studies, because it allows us to gain more insight into the relative contribution of gesture to speech. Some studies found evidence for a decrease in gesture rate when information is shared or repeated (Galati and Brennan, 2014; Jacobs and Garnham, 2007), suggesting that gestures become gradually less important, but others found that it increases (Holler et al., 2011) or that it stays the same (de Ruiter et al., 2012). A smaller number of studies have also considered the form of gestures, and generally these studies found evidence for gestures being smaller and less precise when relating to information in common ground (Galati and Brennan, 2014; Gerwing and Bavelas, 2004; Holler and Stevens, 2007; Vajrabhaya and Pederson, 2013). Gerwing and Bavelas (2004), for example, argue that gestures relating to given information are “sloppier” and more “elliptical”, much like words expressing given information are articulated less clearly.

More recently, Hoetjes et al. (2011, 2015) conducted a large-scale study to gain more insight in gesture behaviour during the production of repeated references, also in view of the mixed results of earlier studies. This was done using a variant of the director–matcher referential communication task (e.g., Clark and Wilkes-Gibbs, 1986; de Ruiter et al., 2012; Holler and Stevens, 2007; Krauss and Weinheimer, 1966), in which speakers were asked to refer to Greebles (Gauthier and Tarr, 1997), which are hard to describe figures with different shapes and protrusions. During the experiment, the director (speaker) described various Greebles to the matcher (addressee), some of which were described multiple times, allowing the authors to compare initial, second and third references. They found, among other things, that the gesture rate (per 100 words) did not differ significantly between the three descriptions. In addition, no reliable qualitative differences in form were found (looking at gesture duration, gesture size, whether the gesture was produced with one hand or with two hands and at the number of repeated strokes). However, in an additional judgment study, they found that gestures produced during initial descriptions were judged to be more “precise” (as defined by Gerwing and Bavelas, 2004) than those produced during repeated descriptions.

2.2. The impact of (negative) feedback

The studies on reduction in referential communication in speech and gesture discussed above all involve situations in which the communication was successful. This was generally the case because the speaker received positive, “go on” feedback, that was either explicit (e.g. via backchannel cues from the addressee) or implicit (e.g. because the addressee selected the right “target”). However, referential communication is not always successful, which an addressee may indicate by responding to an initial description with negative, “go back” feedback. Various studies have revealed that negative feedback signals are marked, in that they are associated with more prosodic effort, for instance because they are realised with a higher pitch, longer duration and more pauses than comparable positive feedback signals (Krahmer et al., 2002; Shimojima et al., 2002). This makes intuitive sense, since it is more important for the speaker to pick up negative than positive feedback from the addressee.

Speakers can respond to negative feedback in various ways, also co-depending on the nature of the feedback. For example, the speaker might repeat the words, but rather than reduce these, she is likely to articulate them with more prosodic effort (louder, higher, etc.), resulting, potentially, in hyper-articulated speech (Lieberman, 1963; Lombard, 1911; Oviatt et al., 1998). In addition, she may reformulate the original utterance and/or add further information to it (Litman et al., 2006). In this study, we investigate whether, and if so, to what extent, a speaker’s gestural behaviour changes as well in response to negative feedback. Given the aforementioned close relationship between speech and gesture, it can be hypothesised that gestures produced during a repeated description following negative feedback are not reduced, but what the precise effect will be on the gesture rate and gesture form is difficult to predict. The outcome does have important implications for theories about speech-gesture production, as it will inform us about the relative importance of the gesture modality during communicative problems.

So far, only a handful of studies have looked at gesture production in response to feedback. Jacobs and Garnham (2007, experiment 2), for example, found an effect of the level of attentiveness of the listener on gesture production. They had participants narrate a comic strip to either an attentive or inattentive confederate listener. The attentive listener was instructed to behave in an attentive manner while each strip was explained, using appropriate verbal and non-verbal (positive) feedback, while the other was instructed to display “inattentive behaviour”. Jacobs and Garnham found that speakers produced more gestures when the listener seemed attentive rather than inattentive. In a somewhat similar vein, Galati and Brennan (2014) point out that speakers take into account verbal and non-verbal addressee feedback, which in turn may shape the speaker's gestures (see also Kuhlen and Brennan, 2010). However, in their study, Galati and Brennan conclude that feedback could not solely account for the way speakers changed their gestures when talking to different addressees (p. 447). While studies such as these indicate that speakers' overall gestural behaviour may be influenced by (lack of) feedback from an addressee, they do not provide insights into the question of how speakers adapt their gestures, both in terms of frequency and form, in response to specific instances of (negative) feedback.

As far as we know, the only study that addresses this question in any detail is Holler and Wilkin (2011). These authors first point to a small number of descriptive studies, describing examples from earlier work which indeed suggest that individual gestures can be adjusted due to feedback from the addressee (Kendon, 2004; Streeck, 1993, 1994). This serves as a starting point for Holler and Wilkin's experimental study, in which they asked participants to retell a fragment from a German television series for children to a confederate addressee who provided scripted feedback at four predetermined points in the narrative. Feedback always took the form of a question, which could either be a request for clarification or confirmation of a detail, or an expression of global non-understanding, asking the speaker to repeat or clarify what was said. Notice that all of these could be classified as "go back" feedback signals, in that they indicate that the addressee requires more information about what the speaker said before. Holler and Wilkin compared utterances before and after feedback, focusing on the gesture rate and the form of gestures. They found that speakers gestured at a numerically slightly higher rate before than after feedback, although this difference was not statistically significant. They then zoomed in on the effects of the four feedback signals separately, and found, again, that for three out of four types of feedback, gesture rate before and after feedback did not significantly differ. The fourth one (seeking confirmation) did lead to a significantly lower gesture rate. Concerning the analysis of gesture form, Holler and Wilkin compared 100 pairs of gestures produced before and after feedback, and found that in the majority (60%) of the cases gestures were likely to be "more communicative" after feedback, which means that they were either larger, more precise (in the sense of Gerwing and Bavelas, 2004), produced in a visually more prominent place or more likely to be displayed from a character perspective (see Holler and Wilkin, 2011, p. 3531, for details).

Holler and Wilkin (2011) point out that their study offers the first insights into how addressee feedback influences gesture production, but they also highlight a number of issues that should be taken up in future research. One concerns the nature of the feedback that was provided; even though feedback was scripted, there was some variation in the behaviour of the confederate, for instance "in terms of whether she used a gesture or not" (Holler and Wilkin, 2011, p. 3534). Given earlier studies on mimicry in gesture production (see e.g., Mol et al., 2012, for an overview and discussion), this could have influenced the gestures produced after feedback. In addition, they point out that it is unclear to what extent their findings can be generalised to different languages (the language they studied was English), other kinds of feedback, and other variables capturing the form of the speaker's gestural behaviour.

2.3. On the role of visibility

Gesture researchers have often used visibility in their experimental designs to get a better understanding of the extent to which gestures are produced for an addressees or whether they are (also) produced for the speaker, i.e., may serve more cognitive needs (see Bavelas and Healing, 2013, for discussion). The general reasoning is that if speakers would produce gestures to further their addressees' understanding, one would expect speakers to produce fewer gestures when addressees cannot see them (see e.g., Alibali et al., 2001, for this argumentation). Indeed, various studies have found that gesture rates decrease when participants cannot see each other (e.g., Alibali et al., 2001; Bavelas et al., 2008). In addition, visibility may also influence the form of the gesture (Bavelas et al., 2008; Gullberg, 2006). For example, Bavelas et al. (2008) found that speakers, describing an elaborate dress on a picture in a mutual visibility condition, used larger gestures, as if they were positioning the dress around themselves, while speakers describing the dress over the telephone tended to produce gestures on the same scale as on the picture.

In line with our previous study on repeated references (Hoetjes et al., 2015), and following many other studies (e.g., Alibali et al., 2001; Bard et al., 2000; Bavelas et al., 2008; de Ruiter et al., 2012; Hoetjes et al., 2014; Holler et al., 2011; Mol et al., 2009), we include visibility as an additional variable in the design of our production experiment (experiment I). We do this in such a way that one group of participants will be able to see each other (mutual visibility), while the other group are prevented to do so using a screen (no visibility). We include visibility in our design for two reasons: first, because it enables comparison with our previous study, on repeated references in successful communication, and, second, to study whether the impact of negative feedback on gesture production, both in terms of gesture rate and in terms of gesture form, is more speaker- or more addressee-oriented.[2]

3. The present study

In this paper, we study the influence of negative feedback on the production of repeated multimodal Dutch referring expressions. For this, we use the same general

[2] Note, however, that manipulating visibility does not necessarily distinguish between speaker and addressee functions, see Bavelas et al., 2008; Holler et al., 2011.

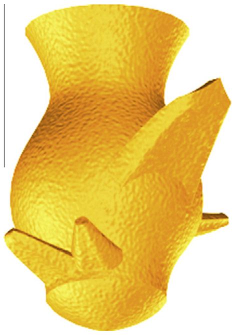

Fig. 1. Example of a Greeble, turned upside down as compared to their presentation in Gauthier and Tarr (1997).

set-up as employed in Hoetjes et al. (2011, 2015), in which speakers, in a director–matcher task, had to refer to hard-to-describe objects with different shapes and protrusions (the aforementioned Greebles). Using the same set-up has two main advantages. Firstly, we know from the aforementioned study that referring to Greebles elicits a substantial number of spontaneous (mostly representational) gestures, both in initial and repeated descriptions. Secondly, and arguably more important, it serves as a kind of baseline, in that it allows us to compare speech-gesture production in successful repeated descriptions with unsuccessful ones, after negative feedback from the addressee.

Feedback (both positive and negative) can come in many variants. Here we opt for a simple variant: after a speaker (the director) has described a target object, the addressee (the matcher, who is a confederate of the experimenter) either selects the correct referent ('go on', which is signalled using a pleasant high ping sound) or (in a limited number of critical, repeated trials) a wrong one ('go back', signalled using a low buzzing sound). The current set-up enabled us to have a large level of control over the negative feedback, which was identical for all participants. In this way we could collect initial (before feedback) and repeated descriptions (after negative feedback) for all speakers for the same targets. This allowed us to study how speakers (which are the unit of analysis in our study, cf. Bavelas and Healing, 2013) adjust their gesture behaviour on the basis of negative feedback.

As mentioned above, following Hoetjes et al. (2015), and many other related studies (e.g., Alibali et al., 2001; Bavelas et al., 2008; de Ruiter et al., 2012; Hoetjes et al., 2014; Holler et al., 2011; Mol et al., 2009), we added visibility as an additional variable to the design, in such a way that one group of participants could see each other during the experiment, while the other group was prevented from doing so by an opaque screen which was placed in between them.

For the critical trials, the initial (pre-feedback) as well as the second and third (post-negative-feedback) descriptions were manually transcribed and the accompanying gestures coded. As motivated above, this allowed us to compare the gesture rate before and after negative feedback across multiple descriptions. In addition, we studied whether the form of the gestures changed as a function of feedback, using the coding scheme employed by Hoetjes et al. (2015), looking at duration and size of the gestures, number of hands involved (one or two) and number of stroke repetitions. Additionally, precision of gestures was assessed using a separate judgment study with naive participants.

By looking at both gesture rate and gesture form before and after negative feedback, we can further our understanding of the role that co-speech gestures play during communication. Gesture rates have often been used in gesture studies, because they inform us about the relative importance of speech and gesture in a multimodal utterance. For example, if gesture rate per word would increase after negative feedback, this would imply that speakers rely more on the gestural modality than on speech in the case of communication problems. In a similar vein, by comparing gesture form before and after negative feedback, we may learn how important gestures are for speakers and how much effort they put into them, and compare this to speech processes after negative feedback. For example, if speakers would produce more precise gestures after negative feedback, this would suggest they put more effort in the gestural part of their utterances. Earlier research on successful communication has often suggested that speech and gesture go "hand-in-hand". In this paper, we ask whether the same pattern can be observed in the case of communication problems, or whether negative feedback has a different impact on gesture and speech production. This offers potentially important information for gesture-speech production models, which aim to explain how speakers produce speech and gesture in tandem (see e.g., Chu and Hagoort, 2014; Hoetjes et al., 2015; Hostetter and Alibali, 2008; Wagner et al., 2014, for recent discussion).

4. Experiment I: Production of gestures before and after negative feedback

4.1. Participants

Participants were 38 undergraduate students (9 male, 29 female, age range 18–30 years old, $M = 21$ years and 7 months), who took part as partial fulfilment of course credits. The participants took part in the experiment in the role of director, and a confederate took part in the role of matcher. This confederate was the same person (female, 20 years old) for all 38 director participants. The participants had no knowledge of, and had not taken part in our previous study on repeated references (Hoetjes et al., 2011, 2015).

4.2. Stimuli

The stimulus materials consisted of picture grids of Greebles[3] (see Fig. 1 for an example Greeble and see

[3] Images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology Carnegie Mellon University, http://www.tarrlab.org/.

Gauthier and Tarr, 1997, for a more detailed description of the Greebles and their properties), which are abstract, small, yellow objects that are hard to describe. The Greebles, which were initially designed to study human face recognition, vary in terms of their gender (“Glip”, “Plok”), their main body shape (“Samar”, “Galli”, “Radok”, “Tasio”), their different types of protrusions (“Boges”, “Quiff”, “Dunth”), and the different shapes and sizes of these protrusions.

We successfully used the same Greeble objects in our previous study on reduction in repeated references (Hoetjes et al., 2011, 2015), and this is the main reason for reusing them in the current study. The Greebles were originally selected because they are quite abstract, and because they only differ from each other with regard to their shape and protrusions. The assumption was that, since speakers would naturally be unfamiliar with the specialised Greeble vocabulary mentioned above (e.g. “Glip”), these shapes and protrusions would have to be described in detail, using both speech and gesture. This way, we could collect repeated object shape descriptions, which were likely to contain repeated gestures illustrating the same Greeble-parts. As in the previous study, the Greebles were turned upside down as compared to the way in which they were presented in Gauthier and Tarr (1997), to make them look less like animate objects (which might cause participants to produce fewer shape descriptions because it would facilitate lexical descriptions such as “angry-looking” or “with the long nose”). We created two picture grids, each containing 16 Greebles. There were 10 trials per picture grid, thus 20 trials in total. In each trial, there was one target object, marked by a red square surrounding it, and 15 distractor objects surrounding the target object (see Fig. 2 for an example of a picture grid). The order in which the directors were presented with the two picture grids was counterbalanced across participants.

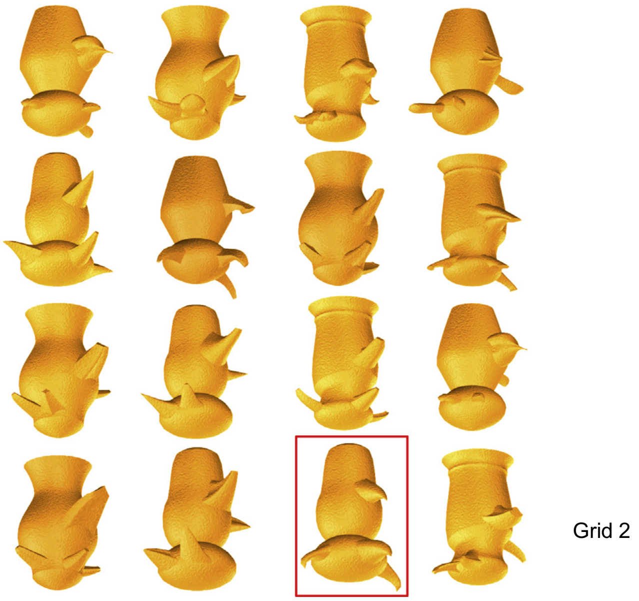

Fig. 2. Example of one of the picture grids (picture grid 2). The target picture of this particular trial is the one in the bottom row, third from left (surrounded by a red square). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The experimental manipulation (and the crucial difference with our previous study, in which we used these same stimuli) was that several Greebles had to be described repeatedly due to apparent communication problems. In each of the picture grids, two Greebles had to be described three times, of which the second and the third description were produced following negative feedback. To make sure that these critical trials did not stand out, an additional seven Greebles per grid had to be described once, and one Greeble had to be described twice (once after negative feedback). These were the filler items. The repeated references to the same object had to be given one straight after the other, when negative feedback provided by the matcher made it clear to the participant that an incorrect object had been chosen (see procedure below). The participants did not know in advance that in some of the trials they would have to take several attempts at describing a picture. This means that the participants thought they had to produce 10 descriptions for each picture grid (one per trial), when in reality they had to produce 15 descriptions for each picture grid. The Greebles that had to be described repeatedly were always preceded and followed by a filler item. To avoid order effects we made sure that the objects that had to be described repeatedly were never in the first or the last trial of the picture grid. We analysed all three descriptions of the objects that had to be described three times (i.e. a total of twelve descriptions for each participant, since four objects had to be described three times).

4.3. Procedure

The experiment consisted of a director–matcher task that was performed in a lab, where the director and the matcher were seated at a table opposite each other (see Fig. 3 for an example of the setup). After entering the lab, the participants (both the director and the confederate matcher) were given written instructions and had the opportunity to ask questions, after which the experiment started. The fact that the matcher was a confederate was to some extent communicated to the director: the director was told that the matcher was someone who had done the experiment before and was helping out because another participant had not shown up. In order to make sure that the director would do her best in providing good descriptions of each target and could not rely on previous experience of the matcher, she was told that the order in which objects were discussed was different for each participant pair (which was not actually the case). The instructions did not mention the use of gesture, so all gesture production was spontaneous.

The director was presented with the trials on a computer screen (which was positioned to her side), and the task for the director was to provide a description of the target

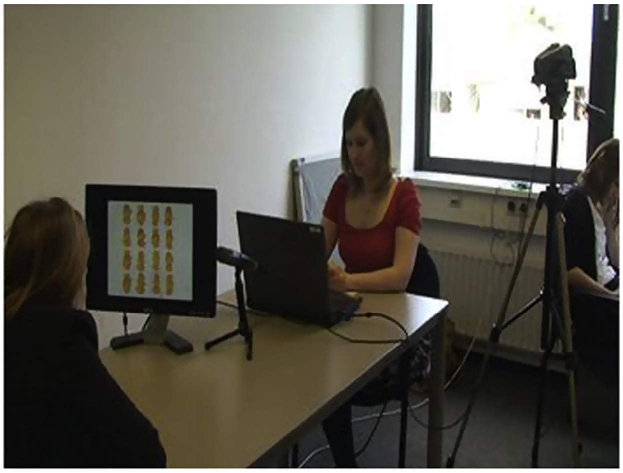

Fig. 3. Example of experimental setup. The director is seen from the back, viewing one of the picture grids. The confederate matcher is seated across from the director, and the experimenter is seated to the side (just visible on the right next to the camera). The director and the matcher can see each other in this example, but for half of the participants a large (around A0 size) opaque screen was placed between the director and the matcher.

object in such a way that it could be distinguished by the matcher from the 15 distractor objects. The director was told that, on the basis of her target description, the matcher picked the object that she thought was being described. After the matcher had picked one of the objects, a sound would tell the director whether the matcher had chosen the correct object or not (a low buzzing sound was played for incorrect object identification and a high ping sound was played for correct object identification). In terms of the coding scheme of Stivers and Enfield (2010), our negative feedback can be seen as an "other-initiation of repair", comparable to the feedback for scene 3 in Holler and Wilkin (2011) and the "What?"/"Sorry"/"huh?" negative feedback used in Healey et al. (2013). When the sound indicating incorrect object identification was played, the director would describe the same target object again, until the matcher had identified the correct object. After this, the director could move on to the next trial. After 10 trials (and a total of 15 descriptions), the director was shown a second picture grid containing 16 new objects, and continued for another 10 trials (i.e. 15 descriptions).

The director was told that the matcher was shown the same objects on her screen (which was positioned in front of her) as on the director's screen, but that these objects were ordered differently for the director and the matcher. It was explained that this meant that the director could not use the object's location in the grid for her target descriptions. In reality however, and unknown to the director, the director and the matcher both viewed the same picture grid and all the matcher had to do was play one of the sounds after the director had given a description of the target object of that particular trial. The participants were debriefed at the end of the experiment, and none of the participants expressed any suspicions concerning the experimental set-up.

The feedback given by the matcher only consisted of the sounds that were played after each trial, although she occasionally added appropriate post-feedback comments such as "hmm, that was the wrong one." The matcher offered no other verbal or non-verbal feedback, and displayed a neutral facial expression throughout the experiment. In addition, the matcher did not interrupt the director, gesture, or ask any questions. This allowed us to collect descriptions before and after negative feedback that were as comparable as possible, to ensure that any effects could be attributed to our manipulation, and not to possible differences in verbal interaction (see Holler and Wilkin, 2009, p. 273 for a similar argument, and Alibali et al., 2001, and Mol et al., 2009, for comparable instructions).

The entire experiment was filmed, with one camera positioned behind the matcher (filming the director) and another camera positioned to the side of the director (filming the entire setup, as in Fig. 3). For half of the participants, a large opaque screen was placed between the director and the matcher, meaning that, in these cases, the director and the matcher could not see each other throughout the entire experiment. Other than that the mutual visibility and no visibility conditions were identical.

4.4. Data analysis

The video recordings were digitised and the recordings showing the director were annotated using the multimodal annotation programme ELAN (Wittenburg et al., 2006). The subsequent (speech and) gesture annotation and data analysis were based on previous research on (reduction in) repeated references, especially the research reported in Hoetjes et al. (2011, 2015).

As a manipulation check, and to enable computation of gesture rate, we first conducted an analysis of the speech. All speech produced within one of the critical references (using the moment when the matcher played one of the sounds as the cut off point) was transcribed orthographically. Hesitations, false starts, repetitions and corrections were all transcribed and included in the word count. Importantly, the distribution of disfluent elements was equal over the various conditions, so that these did not bias the gesture rates reported below. Contractions were counted as single words, but we encountered only one of these in our data ("zo'n" – such a). We analysed the number of words per trial, the duration (in seconds) per trial, and, based on these, we computed the speech rate (in number of words per second) per trial. Based on earlier research we expected the speech rate to go down after negative feedback (Krahmer et al., 2002; Shimojima et al., 2002), and this thus offers a manipulation check.

The gesture annotation was identical to the one employed by Hoetjes et al. (2015), and relied on the gesture phases distinguished by Kendon (1980, 2004), see e.g., also McNeill (1992), Bressem and Ladewig (2011) and Wagner et al. (2014). According to this view, gesture production consists of a number of phases. Starting from a stable, rest position, gesture production begins with a preparation phase, in which the hand moves away from the rest position, after which the stroke occurs, which is usually regarded as the obligatory, main part of the gesture, containing most effort as well as most semantic information. Before or after the stroke, a motionless phase may occur, which is usually referred to as the hold phase. Finally, the hands may return to a rest position during the retraction phase. For the gesture analyses, all stroke phases of all gestures produced in the descriptions of the objects that had to be described three times were selected.[4] The first video frame in which the most effortful movement started was taken as the onset of the stroke, while the offset of the stroke was taken to be the first video frame in which the stroke phase turned into a post-stroke hold, or retraction, phase.

Various authors have emphasised the importance of distinguishing different kinds of gestures during analyses (e.g., Alibali et al., 2001; Bavelas et al., 2008; de Ruiter et al., 2012). Based on McNeill (1992), a distinction can be made between iconic, deictic and beat gestures. Iconic gestures, in our data, are gestures that depict a particular feature of the target object, such as its main shape or the shape of one of the protrusions ("shaped like [this]", where the word 'this' is accompanied by an iconic shape gesture). Deictic gestures are pointing gestures, generally used to indicate a specific location of one of the object's protrusions ("and [here] there is a pointy bit"). Beat gestures consist of simple rhythmic movements without any semantic relation to the speech they accompany. In our previous study, also using Greeble stimuli (Hoetjes et al., 2015), we found that over 95% of the gestures produced by directors were iconics (and, importantly, that figure did not change depending on whether it was an initial or repeated description), making separate analyses for different kinds of gestures impossible. The same applies to the current dataset, in which the affordances of the Greeble stimuli (consisting of distinct shapes and protrusions), resulted in our speakers producing iconic gestures almost exclusively. Therefore we decided, as in Hoetjes et al. (2015), to not distinguish between the different types of gestures in our gesture analyses.

We computed gesture rate per description by dividing the number of gestures by the number of words. For the sake of readability, rates were multiplied with 100, so that the gesture rate can be interpreted as the number of gestures per 100 words. In addition, we analysed several aspects of the form of the gestures. When a director did not produce a gesture in a description, this was treated as a missing value in our analyses on gesture form. The following four aspects of gesture form were taken into account. We measured the duration of the stroke, in seconds. We measured the size of each gesture by coding whether the stroke was produced with a finger (code 1), the hand (code 2), the forearm (code 3) or the entire arm (code 4), with a higher code assuming that the smaller articulators could also be used (e.g. code 3 includes 1 and 2). We coded whether the gesture was produced with one hand or with two hands (resulting in a range from 1 to 2, with e.g. 1.3 indicating that 30% of gestures were two-handed). Finally, we annotated the level of repetition within each gesture by counting the number of repeated strokes. A stroke was considered to be repeated when (nearly) identical strokes followed each other without a retraction phase in between.

To assess annotation reliability, a second annotator, who was not aware of the experimental conditions, coded gesture duration, gesture size, number of hands and number of repeated strokes for a subset of the data, consisting of the first gesture of all participants who produced at least one gesture ($N = 34$ gestures, 2.5% of the data). The annotators agreed on only 44% of cases on gesture duration[5] (Kappa = .042), but on 88% of cases on the size of the gesture (Kappa = .821), 97% of cases on the number of hands that were used (Kappa = .941), and on 73% of cases on the number of repeated strokes (Kappa = .277). The low level of agreement on gesture duration meant that we decided to disregard gesture duration from our further analyses.[6] The other levels of agreement indicate that these annotations were reliable, and range from 'fair', for repeated strokes, to 'almost perfect agreement', according to Landis and Koch's (1977) characterisation. Therefore, we used the first author's annotations for the statistical analysis.

Speech and gesture analyses were conducted for all three reference descriptions of the objects that had to be described three times. The statistical procedure consisted of two repeated measures ANOVAs, one by participants (F_1) and one by items (F_2). On the basis of these, *minF'* was computed (Clark, 1973), so that the results can be generalised over participants and items simultaneously, while keeping the experiment-wise error rate low (Barr et al., 2013, p. 268). The experiment consisted of a 2 × 3 design, with factors Visibility (levels: screen, no screen) and Repetition (levels: initial, second, third), with initial references produced before feedback and second and third references produced after negative feedback from the matcher. We used post hoc analyses and only report where results are significant after correcting for multiple comparisons using the Bonferroni procedure.

4.5. Results

We first discuss effects of repetition and visibility on speech, followed by our main focus: effects of repetition and visibility on gesture rate, and on gesture form.

[4] Given the smaller size of the dataset in this study as compared to Hoetjes et al. (2015), we decided to include all gestures in the detailed analysis, whereas in Hoetjes et al. (2015) only one gesture per description was annotated in detail (even though all were counted and taken into account for analyses of gesture rate).

[5] There was agreement on gesture duration when there was a maximal difference of 5 frames, or 200 ms, between annotators.

[6] Leaving out the analyses for gesture duration did not change the overall picture as presented in the results section since there were no significant effects of repetition or visibility on gesture duration.

4.5.1. Effects on speech

In Table 1, we show the means and standard errors of the dependent speech variables for all three object descriptions. Firstly, inspection of Table 1 reveals that the second references (after negative feedback) were shorter in duration than the initial references, while third references (also following negative feedback) were in turn longer than the second references, but shorter than the initial ones. This effect of repetition was significant, F_1 $(2,72) = 17.17$, $p < .001$, $\eta_p^2 = .323$; F_2 $(2,9) = 7.20$, $p < .05$, $\eta_p^2 = .616$; $minF'$ $(2,18) = 5.07$, $p < .05$. Post hoc Bonferroni analyses showed that all three references differed from each other (all $p < .05$).

Secondly, we found that the second references contained fewer words than the initial references. The third references contained more words than the second references, but fewer than the initial references (see Table 1). This effect of repetition was significant, F_1 $(2,72) = 29.22$, $p < .001$, $\eta_p^2 = .448$; F_2 $(2,9) = 15.91$, $p < .01$, $\eta_p^2 = .780$; $minF'$ $(2,21) = 10.29$, $p < .001$. Post hoc Bonferroni analyses showed that the initial references differed from the second references and from the third references (both $p < .01$). The second and third references did not differ significantly from each other.

Thirdly, as expected, we saw that speech rate (measured in number of words per second) was lower for each following reference (see Table 1). Again, this effect of repetition was significant, F_1 $(2,72) = 30.61$, $p < .001$, $\eta_p^2 = .460$; F_2 $(2,9) = 18.19$, $p < .01$, $\eta_p^2 = .802$; $minF'$ $(2,22) = 11.40$, $p < .001$. Post hoc Bonferroni analyses showed that all references differed from each other (all $p < .01$).

Turning to the effect of visibility on speech, we found that for all three speech variables, a lack of mutual visibility between the director and the matcher caused numbers to go down (see Table 2), although these reductions were only numerical, and not statistically significant. There were no significant interactions between repetition and visibility.

4.5.2. Effects on gesture rate

In Table 3, the means and standard errors of all the dependent variables in gesture for all three object descriptions can be found. Below we discuss them in more detail, starting with number of gestures and gesture rate.

First, we counted the number of gestures per trial. In absolute numbers, fewer gestures were produced in the second references (following negative feedback) than in the initial references (before negative feedback), and more gestures were produced in the third references (also following negative feedback) than in the second references (see Table 3). However, this effect of repetition was only significant over participants and not in the $minF'$ analysis, and hence cannot be considered statistically reliable, F_1 $(2,72) = 4.88$, $p < .05$, $\eta_p^2 = .119$; F_2 $(2,9) = 1.5$, $p = .27$, $\eta_p^2 = .250$; $minF'$ $(2,15) = 1.14$, $p = .34$.

Table 2
Overview of means and standard errors (SE) for dependent variables in speech (duration in seconds, number of words, and speech rate in number of words per second), as a function of Visibility (two levels).

	Visibility (SE)	No visibility (SE)
Duration	35.3 (2.3)	32.5 (2.3)
Number of words	72.5 (5.5)	60.2 (5.5)
Speech rate	2.0 (.06)	1.8 (.06)

Given that the number of words also varies from one description to the next, the gesture rate (number of gestures per 100 words) is more important to analyse, and Table 3 shows that after each instance of negative feedback a higher gesture rate is observed. This effect was significant over participants and items, and marginally significant in $minF'$, F_1 $(2,72) = 7.1$, $p < .01$, $\eta_p^2 = .165$; F_2 $(2,9) = 4.8$, $p < .05$, $\eta_p^2 = .516$; $minF'$ $(2,24) = 2.86$, $p = .077$. Post hoc Bonferroni analyses showed that the gesture rate of the initial references differed from the gesture rate of the third references ($p < .01$).

In Table 4, the means and standard errors of all the dependent gesture variables in the two visibility conditions can be seen. There was a numerical, but not statistically significant, decrease both in the absolute number of gestures, and in gesture rate, when there was no mutual visibility. There were no significant interactions between repetition and visibility on number of gestures or on gesture rate.

4.5.3. Effects on gesture form

When we look at aspects of gesture form (see again Table 3), the statistical analyses showed no significant effect of repetition after negative feedback on gesture size or the number of hands that were used to produce the gestures. We did find a marginally significant effect of repetition on the number of repeated strokes, F_1 $(2,54) = 3.236$, $p = .06$, $\eta_p^2 = .107$; F_2 $(2,9) = 13.645$, $p < .05$, $\eta_p^2 = .752$; $minF'$ $(2,62) = 2.61$, $p = .08$, with an increase for each instance of negative feedback. However, post hoc Bonferroni analyses showed that the three descriptions did not differ significantly from each other.

Turning to the effect of visibility on gesture form (see Table 4), we firstly found that there was no effect of

Table 1
Overview of means and standard errors (SE) for dependent variables in speech (duration in seconds, number of words, and speech rate in number of words per second), as a function of Repetition (three levels). Star indicates significant $minF'$.

	Initial description (SE)	Second description (SE)	Third description (SE)
Duration*	39.7 (2.5)	28.9 (1.6)	33.2 (1.8)
Number of words*	85.0 (6.0)	55.4 (3.4)	58.7 (3.9)
Speech rate*	2.1 (.05)	1.9 (.05)	1.7 (.05)

Table 3
Overview of means and standard errors (SE) for dependent variables in gesture (number of gestures, gesture rate (in number of gestures per 100 words), gesture size (range 1–4), number of hands (range 1–2, with e.g. 1.4. indicating that 40% of gestures were two-handed), and stroke repetition (number of repeated strokes)), as a function of Repetition (three levels). Star indicates marginally significant *minF*′.

	Initial description (SE)	Second description (SE)	Third description (SE)
Number of gestures	3.3 (.49)	2.6 (.38)	3.3 (.52)
Gesture rate*	4.1 (.67)	4.8 (.79)	5.3 (.74)
Gesture size	2.9 (.10)	2.9 (.09)	2.9 (.09)
Number of hands	1.5 (.06)	1.4 (.06)	1.3 (.05)
Stroke repetition*	.33 (.06)	.50 (.10)	.55 (.09)

Table 4
Overview of means and standard errors (SE) for dependent variables in gesture (number of gestures, gesture rate (in number of gestures per 100 words), gesture size (range 1–4), number of hands (range 1–2, with e.g. 1.4. indicating that 40% of gestures were two-handed), and stroke repetition (number of repeated strokes)), as a function of Visibility (two levels). Star indicates significant *minF*′.

	Visibility (SE)	No visibility (SE)
Number of gestures	3.4 (.63)	2.8 (.63)
Gesture rate	5.1 (1.0)	4.3 (1.0)
Gesture size*	3.1 (.10)	2.7 (.11)
Number of hands	1.4 (.07)	1.3 (.07)
Stroke repetition	.41 (.09)	.52 (.10)

visibility on the number of hands or on the number of repeated strokes. There was, however, an effect of visibility on gesture size, F_1 $(1,27) = 9.009$, $p < .01$, $\eta_p^2 = .250$; F_2 $(1,9) = 77.642$, $p < .001$, $\eta_p^2 = .896$; $minF'$ $(1,32) = 8.072$, $p < .01$, with gestures produced when there was a screen between the director and the matcher being smaller than gestures produced when there was no screen between the director and the matcher. There were no significant interactions between repetition and visibility for any of the aspects of gesture form that were analysed.

Summarising the findings of experiment I, we found that references after negative feedback had a lower speech rate and a marginally significant higher gesture rate than initial references. In addition, gestures after negative feedback contained marginally more repeated strokes. When there was no visibility between the director and the matcher, gestures were smaller.

5. Experiment II: Precision judgment

In addition to the gesture measure analyses of the production experiment (experiment I), a separate precision judgment study was run to see whether there might (also) be differences in form between initial gestures and repeated gestures following negative feedback which are more gradual in nature than could be established using the discrete annotations of the data obtained in the production experiment. In this precision judgment experiment, as the name suggests, participants judged the precision of gestures. The setup of this precision judgment experiment, as was the case for the production experiment, closely follows the method used in our previous work on repeated, successful references (see also Hoetjes et al., 2011, 2015).

5.1. Participants

Twenty-nine participants (15 male, 14 female, age range 16–55 years old, $M = 30$ years old), who had not taken part in the production experiment and who had no knowledge of our other previous experiments, took part in the precision judgment experiment, without receiving any form of compensation.

5.2. Stimuli

For the precision judgment experiment, 44 trials were constructed, consisting of 44 pairs of video clips which were selected from the dataset collected in the production experiment. The pairs of video clips consisted of one video clip of a gesture taken from an initial description, and one video clip of a gesture following negative feedback, taken either from a second or third description. We selected all gesture pairs (44) that matched the following criteria. Each pair of gestures was taken from descriptions produced by the same director and both gestures in a pair referred to the same part of the same target object. No more than two gesture pairs produced by one director were used. Also, there had to be an equal distribution between gestures from second and from third descriptions. Of the 44 pairs of video clips, 23 were pairs consisting of one gesture from an initial description and one gesture from a second description, and 21 were pairs consisting of one gesture from an initial description and one from a third description. Finally, we aimed for an equal distribution between visibility conditions. For 19 of the 44 pairs, the gestures were taken from directors who were not able to see the matcher during the production experiment, and the remaining 25 pairs were taken from directors who were able to see each other.

Video clips were presented next to each other in pairs on a computer monitor, and the order in which the clips were presented on the screen was counterbalanced over trials. We presented participants with pairs, and not triads, of gestures, because there were not a sufficient number of directors producing a gesture about the same part of the same object in all three descriptions. Note, however, that in the analyses we did also consider possible differences between gestures from second and third references.

5.3. Procedure

The participants were presented individually with the 44 pairs of video clips. For each pair of video clips, the participants had to judge which of the two gestures they considered to be 'the most precise', where we expected gestures produced during repeated descriptions (i.e. following negative feedback) to be judged more precise than gestures from initial descriptions. No instructions were given with regard to what aspect(s) of the gesture the participants should take into account when making this judgment. The experiment was a forced choice test, presented without sound, and the participants were allowed to watch a video clip more than once if they wanted to. However, they were encouraged to go with their first intuition, and participants made hardly any use of the possibilities for replaying stimuli.[7]

5.4. Data analysis

In each trial, in line with our expectation, a score of one (1) was assigned when the gesture following negative feedback was chosen by the participant to be the most precise, and a score of zero (0) when the participant chose the initial (pre-feedback) gesture to be the most precise. A binomial test was performed to see whether repeated gestures, after negative feedback, were chosen more often than initial gestures to be the most precise one of the two; in addition, a chi square analysis was conducted on the total scores (i.e. number of times that the gesture following negative feedback was chosen to be the most precise), with repetition (pairs of initial and second gestures versus pairs of initial and third gestures) and visibility (mutual visibility versus no mutual visibility) as independent variables.

5.5. Results

Repeated gestures were chosen to be the most precise in 673, or 53%, of 1276 cases, and initial gestures were chosen to be the most precise in 603, or 47%, of cases. This difference from chance level was marginally significant, $p = .053$.

Table 5 shows the distribution of scores for the number of times a gesture following negative feedback was chosen to be the most precise, as a function of repetition (second or third description) and visibility. A chi-square test of independence was conducted to examine the relation between repetition and visibility. We found a significant relation between repetition and visibility, $\chi^2(1) = 15.303$, $p < .001$. A chi-square test of goodness-of-fit showed that there was an equal distribution between repeated gestures from second references and from third references, $\chi^2(1) = 1.618$, $p = .203$. However, there was not an equal distribution between gestures taken from contexts of mutual visibility and gestures taken from contexts without visibility. Gestures following negative feedback which were produced with mutual visibility were chosen more often to be the most precise than gestures following negative feedback which were produced without mutual visibility, $\chi^2(1) = 25.499$, $p < .001$.

Table 5
Distribution of scores (and percentages) for number of times a repeated gesture (i.e. following negative feedback) was chosen to be the most precise, as a function of repetition (i.e. was the repeated gesture from a second or from a third description) and visibility (i.e. was the gesture produced with mutual visibility, or not).

	Second description	Third description	Total
Visibility	216 (32%)	186 (28%)	402 (60%)
No visibility	104 (15%)	167 (25%)	271 (40%)
Total	320 (47%)	353 (53%)	673 (100%)

6. General discussion

When a speaker describes an object or person, the addressee may or may not be able to determine which object or person is referred to. In the former case, when referential communication is successful, the addressee may either explicitly or implicitly indicate this to the speaker using a "go on" feedback cue, and the interaction continues. But in the latter case, when communication is unsuccessful, the addressee will signal this using a more marked "go back" feedback cue (e.g., Krahmer et al., 2002; Shimojima et al., 2002). We know that these negative "go back" cues have an impact on the next utterance of the speaker, which is more likely to be articulated with increased prosodic effort (higher pitch, louder volume, slower speech rate) and to be reformulated or rephrased (e.g., Litman et al., 2006; Oviatt et al., 1998, among many others). But what is the effect of negative, "go back" feedback on gesture production? Only a very limited number of studies have addressed this question so far, of which Holler and Wilkin (2011) is the most explicit, also in stressing that more research in this field is urgently needed.

In this paper, we investigated what happens in gesture when referential communication is unsuccessful. Specifically, we studied repeated references to hard to describe objects (Greebles) with different shapes and protrusions, comparing initial descriptions with descriptions produced after negative feedback. Our experimental method was a variation of earlier work on successful referential communication to these Greebles (Hoetjes et al., 2011, 2015), and we know from these studies that the

[7] For our study on successful repeated references (Hoetjes et al., 2015) we conducted a very similar judgment study, and also experimented with different variants. In particular, in one variant participants were shown the target Greeble along with the two gesture stimuli, and were explained what was intended with one gesture being more "precise" than another ("for example when it provides more information about the shape of the object or when a gesture is more complex", following Gerwing and Bavelas, 2004). Neither of these adaptations influenced the findings of Hoetjes et al. (2015), which is why we opt for the simplest variant (without Greeble picture and explanation of precision) here.

Greebles reliably elicit spontaneous shape gestures, both during initial and repeated references. In general, we rely on a variant of the director–matcher referential communication paradigm (e.g., Clark and Wilkes-Gibbs, 1986; de Ruiter et al., 2012; Holler and Stevens, 2007; Krauss and Weinheimer, 1966), combined with a visibility manipulation such that some participant pairs could see each other (mutual visibility), while others could not. Crucially, in a number of cases, an initial object description was followed by two, consecutive instances of negative, "go back" feedback, indicating that the addressee was not able to match the correct Greeble object to the description of the speaker. As in various earlier studies using the referential communication paradigm (including Hoetjes et al., 2015; Holler and Wilkin, 2011), we look at both the gesture rate (in number of gestures per 100 words), before and after negative feedback, as well as the influence of feedback on the way directors produce gestures. Our analysis of gesture form consisted of both a detailed analysis of 'discrete' properties of the gestures (their size, number of hands involved and number of stroke repetitions), as well as a separate precision judgment experiment, in which naïve judges were asked to determine which of two gestures (one produced before and one after negative feedback) they considered to be the most "precise".

We found, first of all, a marginally significant increase in gesture rate in repeated references following negative feedback, indicating that our speakers started to rely relatively more on the gesture modality when facing referential communication problems. This is different from the pattern that was observed in Hoetjes et al. (2011, 2015), where gesture rate did not change across repeated, successful references. In general, many studies looking at gesture rate in successful communication found that gesture rate remains either the same or is reduced when speakers present information that is repeated or otherwise given in unproblematic interactions (e.g., de Ruiter et al., 2012; Galati and Brennan, 2014; Jacobs and Garnham, 2007, see Hoetjes et al., 2015 for further discussion). Interestingly, the exception is formed by the work of Holler and colleagues, who found that gesture rate increases with repetition in successful communication (Holler et al., 2011), but not after addressee feedback (Holler and Wilkin, 2011). In general, it is difficult to compare gesture rate across different studies (in which speakers are performing different tasks and talk about different things, which in turn may trigger different kinds of gestures), which is one of the main reasons why we opted for re-using the paradigm of our earlier study. In addition, due to the fact that the gesture rate findings of the present study did not reach significance, it is difficult to relate them to previous findings on gesture rate.

However, gesture rate alone is perhaps not sufficiently informative when studying gesture production, a point also made recently by Bavelas and Healing (2013). Gesture form is important as well. Concerning form we found that gestures produced after negative feedback were somewhat more likely to contain repeated strokes (Experiment I) and to be judged as marginally more precise (Experiment II). Again, these patterns are clearly different from what we observed in Hoetjes et al. (2011, 2015), where repeated (successful) references did not contain more strokes (in fact, no changes in 'discrete' gesture form were found), and where gestures from repeated references were *less* likely to be judged as precise than those in initial references.

On balance, the picture that emerges is that references after negative feedback (and in contrast to successful repeated references) showed a tendency towards relying more on gesture (increased gesture rate), and that these gestures showed a tendency towards being produced with more effort (more stroke repetition, more precision), but more research is needed to support this pattern due to the marginality of statistical effects. This pattern of results seems to be consistent with earlier findings on the influence of negative feedback on speech and language (e.g., Litman et al., 2006; Oviatt et al., 1998), and notice, incidentally, that the decrease in speech rate which we observed matches these earlier findings as well.

It is informative to look at some examples of the kind of descriptions that our participants actually produced in this experiment. Example 1 illustrates the increase in gesture rate in the present study.

Example 1. Repeated descriptions of the same object by participant number 36 (in the no visibility condition), in translation from Dutch original, followed by original number of words, number of gestures and gesture rate. The moment at which a gesture was produced is placed between square brackets (dots indicate silence).

Initial description, before feedback

"Uh this one is [again wide in the middle] and thin at the top and the bottom. Uh the circle is a bit average uh in the circle there are three uh points. And at the top there is one and it edges a little [yes it is on the right side but it] also stands a bit to the front. Uh let me think. Uh there are one, two, three, four, four of this shape I think and this is the only one where three [of those] points are at the bottom".

89 words, 3 gestures, gesture rate 3.37

Second description, after negative feedback

"Yes no that is not true I uh am saying it wrong. Yes there are [two where three] are uh let's have a good look, yes there are two which have three of those uh points at the bottom, only with that one it is again uh uh [it again has the shape of an uh] [...] of such a [yes] [such a handle] of something and the others are a bit more pointy".

71 words, 5 gestures, gesture rate 7.04

Third description, after repeated negative feedback

"Uh let's see. The difference still with those others is that that point at the top that that one does not have those [uh uh] how do you call that [that sort of detail in it], has [detail in it].

37 words, 3 gestures, gesture rate 8.11

Inspection of this example confirms, first of all, that talking about Greebles is hard, but it also illustrates what causes the increase in gesture rate that we observed. While speakers use fewer words in descriptions after negative feedback, they continue to rely on shape gestures, since these express the most distinguishing properties of the target Greeble.

Fig. 4 illustrates increased gesture precision after negative feedback, as compared to before feedback was given. Notice that the gesture after negative feedback is produced at a higher location and shows a larger displacement of the speaker's hands than the gesture before feedback, consistent with the notion that after negative feedback, gestures are produced with more effort.

Since it was used in many relevant earlier studies (most notably for our current purposes, in Hoetjes et al., 2015, but also, for instance, in Alibali et al., 2001; Bard et al., 2000; Bavelas et al., 2008; de Ruiter et al., 2012; Holler et al., 2011; Mol et al., 2009), we included mutual visibility as a factor in our current experiments as well. As in Hoetjes et al. (2015) and many other studies, we found that gestures produced without visibility were smaller than those produced when there was mutual visibility between director and matcher (see Fig. 5). Perhaps more interestingly, we found in the judgment study that when there was mutual visibility, gestures produced after negative feedback were somewhat more likely to be judged as precise than initial, pre-feedback gestures. This suggests that our directors put more effort in their post-feedback gestures when these could be seen by their addressee, which in turn might imply that these gestures were communicatively intended. Notice that this is also in accordance with Holler and Wilkin's (2011) finding that gestures after feedback were "more communicative".

As mentioned before, not many studies have investigated the effect of feedback on gesture production, especially not with regard to the question of how speakers adapt the frequency and form of their gestures. One notable exception, as discussed, is the study on the effect of addressee feedback on gesture production by Holler and Wilkin (2011). As we have seen, our findings, in particular those related to gesture form, appear to be consistent with theirs; after (negative) feedback, gestures appear to be more effortful and communicative. It is interesting to observe that this convergence of results is obtained despite differences in experimental set-up which were partly motivated from their suggestions for further research (Holler and Wilkin, 2011, p. 3534): different kinds of feedback (even though all, as said, are intuitive "go back" signals) which were administered in a different way, different gesture analyses, and different languages. Additionally, while in the current study we compared initial references with *two* instances following negative feedback, Holler and Wilkin (2011) offered at most *one* instance of negative feedback for an individual referent or event. Moreover, we added a visibility manipulation, as well as a separate gesture precision judgment experiment, adding further evidence that gestures after (negative) feedback are somewhat more precise, in particular when they were visible for the addressee.

Various avenues for future research remain. We opted for artificial negative feedback (a low buzzing sound), identical for all participants, administered by a matcher who otherwise remained neutral in her verbal and non-verbal feedback, and did not further interact with the directors. This kind of high level feedback, which may be likened to a "huh?" or "sorry?", indicating that the previous utterance from the director was not successful, has been used before and has the advantage for current purposes that it allowed us to collect comparable descriptions, including gestures, before and after negative feedback, to see how speakers (our unit of analysis, cf. Bavelas and Healing, 2013) adapt their gestures after negative feedback. However, we cannot rule out the possibility that occasionally the matcher did produce some unintentional nonverbal feedback, which the director could subsequently have picked up. In addition, the matcher timed the occurrence of the negative feedback to produce it at the contextually appropriate time, but this also may have introduced some timing differences across trials. In follow up research, it would be important to see whether the findings obtained in the current, controlled set-up, generalise to more natural situations. Ideally, this would involve spontaneous

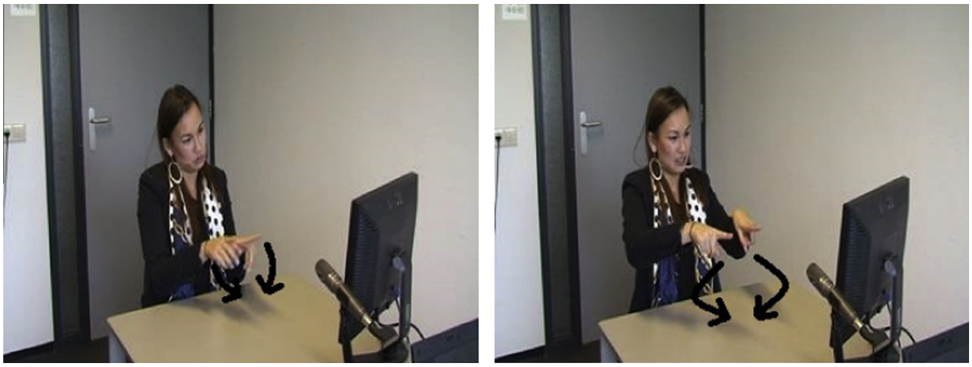

Fig. 4. Example of a pair of gestures produced about the same object by the same participant (in the visibility condition), illustrating gesture precision. The gesture on the left is an initial gesture, produced before feedback, the gesture on the right is a gesture produced after negative feedback, which was judged to be more precise. Arrows indicate path and direction of each gesture.

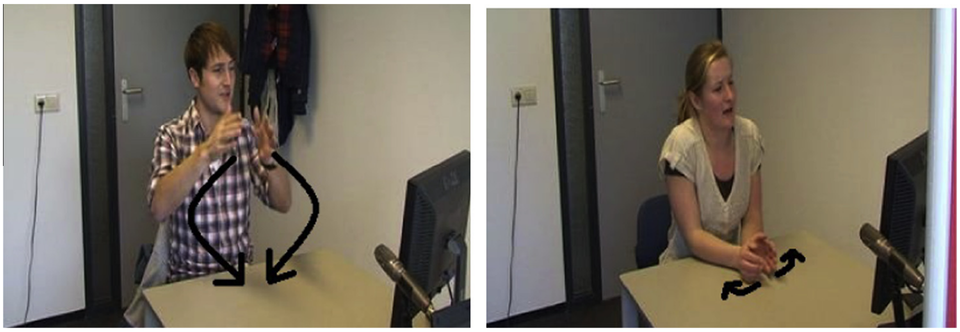

Fig. 5. Example of gestures produced in context of visibility (on the left), and in context without mutual visibility (on the right, part of the opaque screen is just visible), illustrating gesture size. Arrows indicate path and direction of each gesture. In the gesture on the left, the entire arm is moving, whereas in the gesture on the right only the hands are moving.

interactions between pairs of naïve participants, rather than between participants and a confederate, to rule out any undesired experimental side effects of using the latter (cf. Kuhlen and Brennan, 2013). This could involve, for example, communication about Greebles as well, in which miscommunications (of various kinds) may occur in a more natural way.

It is to be expected that, in such a setting, different kinds of feedback and, related, different kinds of interaction, could lead to different gesture patterns. Imagine, just by way of example, that a director describes (in speech and gesture) a Greeble from the Radok family, with a cylindrical main shape. In the current experiment, such an utterance would be followed by general negative feedback. But now consider a different, more specific form of negative feedback, in which the matcher asks (incorrectly) "you mean the one with a vase shape?" (i.e., a "Galli"), indicating this vase shape using a gesture. This "go back" signal from the matcher would likely also initiate a repair from the director ("No, cylindrical."), and may result in a pair of spontaneous cylindrical gestures before and after feedback (comparable to the pairs collected with the current paradigm, except that the negative feedback was specific rather than general). It would be very interesting to compare such pairs (assuming they can be collected in sufficiently large numbers) using a more natural variant of the methodology of the current paper, where we predict that, crucially, the post feedback gestures will be realised with more effort (e.g., more repeated strokes along a virtual cylinder) and are more likely to be judged as precise compared to the pre-feedback counterpart, perhaps to a larger extent than found in the current study.

Related, it would be interesting to see whether our current findings can be generalised to other types of gesture. In the present study, almost all gestures that were produced by directors were representational, and specifically iconic, ones. This was to be expected, since the stimuli were selected on the basis of their differences in shape and protrusions and thus afforded in particular the production of iconic gestures. A question is whether an increase in gesture rate and gesture form similar to what we found in the present study could be seen if the gestures in question were, for example, deictic or beat gestures (or metaphoric gestures or emblems, for that matter). There has been at least one study investigating deictic gestures in repeated references (de Ruiter et al., 2012), but this study did not focus on miscommunication, and studied gesture rate, and not gesture form. It would be interesting to include negative feedback in that type of study, either in the controlled manner ("beep!") of the current study, or the less-controlled, but more natural alternative just sketched ("You mean this one?", while pointing to an incorrect object).

Finally, a last aspect that could be studied in future work concerns the gesture rate, where our findings (marginally significant increase in gesture rate after negative feedback) do not match those of Holler and Wilkin (2011) (no increase after feedback). As we discussed in detail elsewhere (Hoetjes et al., 2015), the study of gesture rate (as a dependent variable in different kinds of studies) has given rise to a complex pattern of results, which may partly be due to different ways in which gesture rates have been computed in the past. In future research, it would seem to be important to more systematically compare different ways of computing gesture rates, to get a better understanding of what these rates may tell us, and why the results can differ from one study to the next. In addition, as we already pointed out above, it becomes increasingly important to combine analyses of gesture rate with analyses of gesture form, to get a better understanding of the gestures that speakers produce.

7. Conclusion

In this study, we asked what happens in gesture when referential communication is unsuccessful. We conducted a director–matcher task in which directors had to produce repeated references about the same object after negative feedback which indicated that communication was unsuccessful. We found that after negative feedback, there was a marginally significant increase in gesture rate and gestures were produced with somewhat more repeated strokes (also marginally significant in *minF′*). In addition, a separate precision judgment test showed that after negative feedback, gestures were somewhat more likely to be rated as most precise, compared to gestures produced before negative feedback was given. Taken together, we suggest that this means that when communication was unsuccessful in our task, speakers showed a tendency towards relying more on gesture, and the gestures they produced trended towards being more effortful. In addition, the visibility manipulation suggests that our directors put more effort in their gestures when these could be seen by the addressee, which in turn might imply that these particular gestures were communicatively intended. All in all, the picture that emerges is rather different from our earlier reduction findings for successful repeated references (Hoetjes et al., 2011, 2015); when communication is successful and information becomes more predictable, speakers can permit themselves to put less effort in their repeated references, both in speech (e.g., less clear articulation, fewer words) and in gesture (e.g., less precision). When communication is not successful, speakers have to make an extra effort, in an attempt to restore communicative success. We already knew that this increased effort has an impact on speech; the current paper suggests that it has a comparable effect on gesture production as well.

Acknowledgments

We would like to thank Elsa Jonkers and Kristel Bartels for their help in collecting the data for both experiments, Elsa Jonkers for additional help in transcribing and annotating the data, Nick Wood for technical support, and Martijn Goudbeek, Judith Holler and an anonymous reviewer for helpful comments and suggestions. An earlier version of this study was presented at the 36th Annual Conference of the Cognitive Science Society, Quebec (2014). We received financial support from The Netherlands Organization for Scientific Research, via a Vici grant (NWO grant 27770007), which is gratefully acknowledged.

References

Alibali, M., Heath, D.C., Myers, H.J., 2001. Effects of visibility between speaker and listener on gesture production: some gestures are meant to be seen. J. Mem. Lang. 44, 169–188.

Aylett, M., Turk, A., 2004. The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. Lang. Speech 47 (1), 31–56.

Bard, E.G., Anderson, A.H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., Newlands, A., 2000. Controlling the intelligibility of referring expressions in dialogue. J. Mem. Lang. 42, 1–22.

Barr, D.J., Levy, R., Scheepers, C., Tily, H., 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. J. Mem. Lang. 68, 255–278.

Bavelas, J., Healing, S., 2013. Reconciling the effects of mutual visibility on gesturing: a review. Gesture 13 (1), 63–92.

Bavelas, J., Gerwing, J., Sutton, C., Prevost, D., 2008. Gesturing on the telephone: independent effects of dialogue and visibility. J. Mem. Lang. 58, 495–520.

Brennan, S., Clark, H., 1996. Conceptual pacts and lexical choice in conversation. J. Exp. Psychol. 22 (6), 1482–1493.

Bressem, J., Ladewig, S.H., 2011. Rethinking gesture phases: articulatory features of gestural movement? Semiotica 184 (1/4), 53–91.

Brown, G., 1983. Prosodic structure and the given/new distinction. In: Cutler, A., Ladd, D.R. (Eds.), Prosody: Models and Measurements. Springer-Verlag, New York, pp. 67–78.

Chu, M., Hagoort, P., 2014. Synchronization of speech and gesture: evidence for interaction in action. J. Exp. Psychol. Gen. 143 (4), 1726–1741.

Clark, H., 1973. The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. J. Verb. Learn. Verb. Behav. 12, 335–359.

Clark, H., Schaeffer, E.F., 1989. Contributing to discourse. Cogn. Sci. 13, 259–294.

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

Cohen, A., Harrison, R.P., 1973. Intentionality in the use of hand illustrators in face-to-face communication situations. J. Pers. Soc. Psychol. 28 (2), 276–279.

de Ruiter, J.P., Bangerter, A., Dings, P., 2012. The interplay between gesture and speech in the production of referring expressions: investigating the trade-off hypothesis. Top. Cogn. Sci. 4 (2), 232–248.

Fowler, C.A., Housum, J., 1987. Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction. J. Mem. Lang. 26 (5), 489–504.

Galati, A., Brennan, S., 2014. Speakers adapt gestures to addressees' knowledge: implications for models of co-speech gesture. Lang. Cogn. Neurosci. 29 (4), 435–451.

Gauthier, I., Tarr, M., 1997. Becoming a "Greeble" expert: exploring mechanisms for face recognition. Vision. Res. 37, 1673–1682.

Gerwing, J., Bavelas, J., 2004. Linguistic influences on gesture's form. Gesture 4, 157–195.

Gullberg, M., 2006. Handling discourse: gestures, reference tracking, and communication strategies in early L2. Lang. Learn. 56 (1), 155–196.

Healey, P., Mills, G.J., Eshgi, A., 2013. Making things worse to make them better: the role of negative evidence in the coordination of referring expressions. In: Paper Presented at the Pre-Cogsci 2013: Production of Referring Expressions: Bridging the Gap Between Cognitive and Computational Approaches to Reference, Berlin.

Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., Swerts, M., 2011. GREEBLES Greeble greeb. On reduction in speech and gesture in repeated references. In: Carlson, L., Hoelscher, C., Shipley, T.F. (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society. Cognitive Science Society, Boston, pp. 3250–3255.

Hoetjes, M., Krahmer, E., Swerts, M., 2014. Does our speech change when we cannot gesture? Speech Commun. (57), 257–267

Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., Swerts, M., 2015. Reduction in gesture during the production of repeated references. J. Mem. Lang. 79–80, 1–17.

Holler, J., Stevens, R., 2007. The effect of common ground on how speakers use gesture and speech to represent size information. J. Lang. Soc. Psychol. 26 (1), 4–27.

Holler, J., Wilkin, K., 2009. Communicating common ground: how mutually shared knowledge influences speech and gesture in a narrative task. Lang. Cogn. Proc. 24 (2), 267–289.

Holler, J., Wilkin, K., 2011. An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses. J. Pragmatics 43, 3522–3536.

Holler, J., Tutton, M., Wilkin, K., 2011. Co-speech gestures in the process of meaning coordination. In: Paper Presented at the 2nd GESPIN – Gesture & Speech in Interaction Conference, Bielefeld.

Hostetter, A.B., Alibali, M., 2008. Visible embodiment: gestures as simulated action. Psychon. Bull. Rev. 15 (3), 495–514.

Jacobs, N., Garnham, A., 2007. The role of conversational hand gestures in a narrative task. J. Mem. Lang. 56, 291–303.

Kendon, A., 1972. Some relationships between body motion and speech. In: Seigman, A.W., Pope, B. (Eds.), Studies in Dyadic Communication. Pergamon Press, New York, pp. 177–216.

Kendon, A., 1980. Gesture and speech: two aspects of the process of utterance. In: Key, M.R. (Ed.), Nonverbal Communication and Language. Mouton, The Hague, pp. 207–227.

Kendon, A., 2000. Language and gesture: unity or duality? In: McNeill, D. (Ed.), Language and Gesture. Cambridge University Press, Cambridge, pp. 47–63.

Kendon, A., 2004. Gesture. Visible Action as Utterance. Cambridge University Press, Cambridge.

Kita, S., Özyürek, A., 2003. What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. J. Mem. Lang. 48, 16–32.

Krahmer, E., Swerts, M., Theune, M., Weegels, M., 2002. The dual of denial: two uses of disconfirmations in dialogue and their prosodic correlates. Speech Commun. 36, 133–145.

Krauss, R.M., Weinheimer, S., 1966. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. J. Pers. Soc. Psychol. 4, 343–346.

Kuhlen, A., Brennan, S., 2010. Anticipating distracted addressees: how speakers' expectations and addressees' feedback influence storytelling. Discourse Process. 47 (7), 567–587.

Kuhlen, A., Brennan, S., 2013. Language in dialogue: when confederates might be hazardous to your data. Psychon. Bull. Rev. 20 (1), 54–72.

Lam, T.Q., Watson, D.G., 2010. Repetition is easy: why repeated referents have reduced prominence. Memory Cogn. 38 (8), 1137–1146.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (159–174).

Levy, E., McNeill, D., 1992. Speech, gesture and discourse. Discourse Process. 15, 277–301.

Lieberman, P., 1963. Some effects of semantic and grammatical context on the production and perception of speech. Lang. Speech 6 (3), 172–187.

Litman, D., Swerts, M., Hirschberg, J., 2006. Characterizing and predicting corrections in spoken dialogue systems. Comput. Linguist. 32, 417–438.

Lombard, E., 1911. Le signe de l'elevation de la voix. Ann. Maladies Oreille Larynx, Nez. Pharynx 37, 101–119.

Masson-Carro, I., Goudbeek, M., Krahmer, E., 2014. On the automaticity of reduction in dialogue: cognitive load and repeated multimodal references. In: Bello, P., Guarini, M., McShane, M., Scassellati, B. (Eds.), Proceedings of the 36th Annual Meeting of the Cognitive Science Society. Cognitive Science Society, Quebec City.

McNeill, D., 1985. So you think gestures are nonverbal? Psychol. Rev. 92, 350–371.

McNeill, D., 1992. Hand and Mind. What Gestures Reveal About Thought. University of Chicago Press, Chicago.

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

Mol, L., Krahmer, E., Maes, A., Swerts, M., 2009. The communicative import of gestures. Evidence from a comparative analysis of human–human and human–machine interactions. Gesture 9 (1), 97–126.

Mol, L., Krahmer, E., Maes, A., Swerts, M., 2012. Adaptation in gesture: converging hands or converging minds? J. Mem. Lang. 66, 249–264.

Oviatt, S., MacEachern, M., Levow, G.-A., 1998. Predicting hyperarticulate speech during human–computer error resolution. Speech Commun. 24, 87–110.

Shimojima, A., Katagiri, Y., Koiso, H., Swerts, M., 2002. Informational and dialogue-coordinating functions of prosodic features of Japanese echoic responses. Speech Commun. 36, 113–132.

So, W.C., Kita, S., Goldin-Meadow, S., 2009. Using the hands to identify who does what to whom: gesture and speech go hand-in-hand. Cogn. Sci. 33, 115–125.

Stivers, T., Enfield, N., 2010. A coding scheme for question–response sequences in conversation. J. Pragmatics 42, 2620–2626.

Streeck, J., 1993. Gesture as communication I: its coordination with gaze and speech. Commun. Monogr. 60 (4), 275–299.

Streeck, J., 1994. Gesture as communication II: the audience as co-author. Res. Lang. Soc. Interact. 27 (3), 239–267.

Traum, D.R., 1994. A Computational Theory of Grounding in Natural Language Conversation. Unpublished PhD Dissertation. University of Rochester.

Vajrabhaya, P., Pederson, E., 2013. Repetition vs. listener accommodation: a case study of co-speech gesture in retellings. In: Paper Presented at the New Ways of Analyzing Variation 42 Conference, Pittsburgh.

Wagner, P., Malisz, Z., Kopp, S., 2014. Gesture and speech in interaction: an overview. Speech Commun. 57, 209–232.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H., 2006. ELAN: a professional framework for multimodality research. In: Paper Presented at the LREC 2006, Fifth International Conference on Language Resources and Evaluation, Genoa, Italy.