Not so dependent after all: functional perception of speakers’ gestures with and without speech

Kasper Kok 1, Kirsten Bergmann2, Stefan Kopp2

1 Faculty of Humanities, VU University, Amsterdam, the Netherlands
2 CITEC, Faculty of Technology, Bielefeld University, Bielefeld, Germany

ki.kok@vu.nl, kirsten.bergmann@uni-bielefeld.de, skopp@techfak.uni-bielefeld.de

Abstract

Speakers’ spontaneous gestures are traditionally thought of as a dependent system, their meaning relying heavily on what is expressed verbally. Nonetheless, studies using muted videos as stimuli consistently report that gestures have some degree of communicative import in the absence of speech. Here, we argue that the dependence-autonomy question can be advanced by adopting a functional linguistic perspective. We ask whether access to speech is necessary to understand what kind of semantic or pragmatic function a gesture performs. Based on a large-scale web-based perception study, we report that when the notion of meaning is operationalized on a functional level, the dependence of gesture on speech largely diminishes.

Index Terms: gesture, multimodality, functional linguistics

1. Introduction

Whereas gestures undoubtedly have communicative value [1, 2], it is commonly assumed that their meaning is underspecified and dependent on the accompanying speech. As noted by McNeill [3,4], the same object or situation can be referred to by gestures with different syntactic and semantic characteristics in different contexts. This context-dependence, according to McNeill, “contrasts with the stability of lexical forms, which present modifications of the core meaning in all contexts but do not give up their basic form in this process.” If no such basic and context-independent forms exist in gestural expression, it can moreover be expected that the meaning of a gesture cannot be determined when the access to the accompanied speech is denied. This hypothesis has been tested in a variety of studies where participants’ performance on comprehension tasks was compared in conditions with and without access to speech [4–7].

Feyereisen et al. [4] tested whether participants were able to qualify the gestures of speakers in naturalistic discourse as either iconic or metaphoric, when videos were heard with or without sound. The authors reported that raters were substantially more accurate when the auditory channel was available, but performance was still above chance level when access to the speech was denied. In a second experiment, it was found that when video clips were presented without access to the audio channel, subjects were generally unable to guess the original utterance from a fixed list of options. However, a high degree of consistency was found in participants’ responses to this task. According to the authors, this hints at the existence of intrinsically meaningful qualities of the gestures in their stimulus set.

A similar paradigm was used by Krauss et al. [5], who presented participants with audio-miuted video clips and asked them to guess which of two words was the gesture’s ‘lexical affiliate’ (a term used after Schegloff [8], referring to “the word or words deemed to correspond most closely to a gesture in meaning”). Raters’ performance on this task was far from perfect, yet significantly above chance level. In a follow-up experiment, participants were instructed to assign semantic categories (action, location, object name, or description) to a set of gestures in conditions with and without sound. It was found that when the speech was available, the classifications very closely reflected the semantic content of the accompanied speech. In the absence of the verbal channel, the judgments were not random: gestures were often assigned the same semantic category as their lexical affiliate, to which participants had no access. Given that the presence or absence of speech was consistently found to have some degree of predictable value, the authors conclude that although gestures can convey some information, they are not richly informative, and the information they convey is largely redundant with speech [5,7].

In accordance with this conclusion, Hadar and Pinchas-Zamir [6] argue that the meaning of gestures is best understood in terms of different levels of semantic specificity: some gestures have a specific, lexeme-like meaning, whereas others convey meaning in a “vague and tentative” fashion. Taking this notion as their point of departure, two experiments were carried out where participants had to select a word from a list that related most closely to a given gesture. Participants less often chose the word that had been coded as the lexical affiliate of the gesture when the speech was muted than when the speech or a transcription of it was present. Among the ‘erroneous’ responses, moreover, visually or semantically related distractor words were chosen more often than unrelated distractors. Based on this graded effect, the authors claim that gestures are on a cline of degrees of semantic specificity.

A recent study by Kibrik & Molchanova [7] used a more contextualized task to investigate the relative dependence of speech, gesture and prosody. Participants watched segments of movies or videotaped conversations in conditions with audio only, video only, prosody only, or combinations of these. They then answered a set of multiple choice questions about the content of the movie clips (e.g. “What does Tamara offer Masha before the beginning of the conversation?”). In line with previous findings, it was found that although participants were more accurate when speech was available, a substantial subset of the questions was answered correctly in the absence of speech as well. On the basis of this result, the authors argue that whereas speech might be the ‘leading channel’ of information conveyance, gestures (and prosody) do carry some degree of independent semantic load.

All four of these papers report a rather ambivalent relation between speech and gestures: on the one hand, gestures alone are not as informative as speech-gesture combinations. On the other hand, gestures by themselves provide sufficient cues for participants to score well above chance level on various types of comprehension tasks. The authors agree that this ambivalence reflects the fact that gestures are semantically ‘underspecified’: they carry some intrinsic meaning, but only on a rather schematic level (cf. [9, 10]). In accord with this claim, it has been demonstrated that the degree of semantic congruence between speech and gesture influences processing latency [11]. It remains a relatively open question, however, how the level of schematicity present in co-speech gestures is best characterized. What types of meaning are associated with gestural forms irrespective of speech, and on what level of abstraction?

Getting a grip on this question requires a comprehension task of a less specific character than those used in the research discussed above. Previous studies commonly assessed utterance comprehension using very concrete and detailed questions (e.g., “what is the lexical affiliate of this gesture?” or “what did X say to Y?”). Thus, the notion of meaning is often simply conceived as reference to some specific object or situation – a view that has been under heavy dispute in contemporary cognitive-functional linguistics. An alternative perspective is to characterize meaning in terms of the functional contribution of an expressive unit to the ongoing discourse. According to recent work, this view provides an appropriate level of abstraction for studying semantic and pragmatic interfaces between speech and gesture [12-14]. In terms of experimental design, this entails that instead of asking participants to identify what object or event a gesture refers to exactly, one needs to ask whether access to speech is needed to infer what kind of semantic or pragmatic function a given gesture performs.
In this paper, we examine the question of speech-gesture dependence, adopting and implementing a functionally inspired perspective. In particular, we examine four prominent functions of gestures: object reference, attribution of a static property (e.g., depicting the shape or size of an object), attribution of a dynamic property (e.g., depicting a process of movement), and meta-communicative signalling (e.g., indicating difficulty in speech production). The perception of these functions by naïve observers is investigated in an online perception study with the presence or absence of speech as the critical manipulation.

2. Methods

The research procedures are akin to those described in [15], supplemented with an audio-only condition. In a web-based study, participants watched video fragments of direction-giving discourse recorded in a relatively natural setting. They were asked to indicate their interpretation of the gestures in these videos by filling in a controlled survey. This survey consisted of a number of statements (e.g. The hands of the speaker refer to an object or person) to which participants assigned an agreement score on a 7-point Likert-scale. In a between-subject design, the same videos were presented with the sound on (Sound-On condition) and with the sound muted (Sound-Off condition). The substantial data set yielded by this design (449 gestures, 16 raters per gesture) allows for a detailed assessment of the degree to which the functional interpretation of speakers’ gestures depends on observers’ access to the co-expressed speech.

2.1. Stimuli

The video snippets that served as stimuli were fragments of the Bielefeld Speech and Gesture Alignment corpus (SaGA, [16]). This corpus consists of German-spoken dialogues where one participant gives directions to another participant after having taken a tour through a virtual town. The stimulus set used for this study contained 173 videos and a total of 449 individually marked gesture strokes. To promote ecological validity, all of the gestures of the route-giver were treated as potential stimuli. That is, we did not filter out any gestures based on a priori interest. The video fragments that were used as stimuli were taken from five different dialogues. The start and end points of these clips were moments when the hands of the speaker were in a neutral position and/or the speech was paused. Thus, the stimuli comprised relatively isolated discourse units.

Most of the stimuli contained more than one gesture stroke (isolated, single-stroke gestures are rare in naturalistic conversation). Therefore, numbers were edited into the video clips in coincidence with the individual stroke phases, so that participants could be instructed to pay attention only to one specific stroke at a time (see Figure 1). In order to prevent the video fragments from being too long, gesture sequences with more than six consecutive strokes were discarded from the stimulus set. For all videos, two versions were created: one was simply a fragment of the original corpus, and one was a version of this fragment where the sound was muted.

2.2. Participants

Participants were recruited and reimbursed via the online research platform Crowdflower. In order to ensure reliable results of the data, the following performance thresholds were implemented. Participants’ responses were included only if they (1) had passed 30% of the test questions correctly presented to assess their attention (see 2.3); (2) had taken more than seven minutes to complete the survey (mean completion time was 21.52 minutes); (3) had sufficient variance in their data (those participants who consistently gave only one or two unique answers to all questions were excluded). After having applied these procedures, a total of 347 participants remained, resulting in 16 judges for all stimuli. The participants were aged 16 to 69 (M=37.3, σ=12.1) and 129 were female. All were present in Germany at the time of participation (according to their IP-address) and reported to have full proficiency of the German language.

2.3. Procedure

Participants were assigned either to the Sound-on or the Sound-off version of the experiment. Before the survey was presented, their availability over well-functioning video and audio equipment was verified by means of a digital ‘captcha’, where visually and auditorily presented words had to be typed in, in order to proceed. Participants who were able to do so were presented with the following instructions:

In this survey you will be asked to answer questions on short video segments. The questions concern the relationship between the speaker’s gestures and what he/she says. Every list of questions concerns one single gesture. When multiple gestures occur in a video segment, they will be marked with numbers.

Please read the questions carefully and answer on the basis of your own perception. Be aware that for some videos none of the options is applicable, while for other videos, multiple options may apply. In the video segments, someone explains a route through a city to his/her interlocutor. You can play the video as often as you want.

Hereafter, a page was displayed with an embedded video on top and a list of statements below. These statements were inspired by the levels of semantic and pragmatic analysis recognized in functional linguistics (cf. [12] for details), but formulated in a simplified way to be accessible to naïve participants. In the current study, we focus on a subset of four possible functions of gestures. Three of these correspond roughly to three prevalent semantic functions of linguistic constituents: gestures’ ability to refer to entities in the world (as might also be realized in speech by nouns or noun phrases like ‘the cup’); gestures’ ability to attribute a static feature to some entity (as might be realized in speech by an adjective like ‘big’); and gestures’ ability to attribute a dynamic property to some entity (as might be realized in speech by a verbal phrase like “is spinning”). The fourth question concerns what one might call a ‘meta-communicative’ function of gesture: the signalling of difficulty with word retrieval. Note that although these functions are examined as individual variables in the current paper, they may be intercorrelated; a single gesture may for instance perform referential as well as attributive functions [15]. Table 1 lists the questions/statements corresponding to the four functions of interest, as well as the labels that will henceforth be used for referring to them.

![](http://www.crowdflower.com)

**Table 1. Survey questions**

<table>
<thead>
<tr>
<th>Question label</th>
<th>English translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Refer-to-Object</td>
<td>The hands of the speaker refer to a specific object or person.</td>
</tr>
<tr>
<td>Depict-Shape</td>
<td>The hands of the speaker depict the size or shape of an object or person.</td>
</tr>
<tr>
<td>Depict-Movement</td>
<td>The hands of the speaker depict the movement of an object or person.</td>
</tr>
<tr>
<td>Word-Search</td>
<td>The hands of the speaker show that he/she has difficulty finding the right words.</td>
</tr>
</tbody>
</table>
Participants’ task was simply to indicate whether they thought each of the statements applied to the gesture of interest (referred to via the number that appeared in the screen.) These judgments were given on a 7-point Likert scale, ranging from ‘certainly does not apply’ to ‘certainly applies’.

The experiment was preceded by a practice trial, so that participants could get accustomed to the task. Subsequently, twenty video clips were presented, one at a time, which were randomly sampled from the total stimulus set. Thus, every participant was exposed with a different subset of video snippets. Because all analyses of interest are item-based, we allowed participants to take part multiple times, with a maximum of five. All cases where the same video had coincidentally been assigned to a participant twice were automatically filtered out from the data set. For each video stimulus, the order of the survey questions was randomized. Test questions that served as diagnostic for participants’ attention were added to the list of survey question (in an unmarked, randomly determined location) for one out of every four questions. These asked to simply tick one of the boxes in the Likert-scale (e.g., the third one from the left).

3. Results

Two types of analysis were conducted to gain insights into the role of speech with respect to the functional interpretation of the gestures. To find out whether participants’ perception of the gesture’s functions varied with the presence or absence of speech across the board, we compared the frequency distributions of the mean ratings on all videos across conditions. Second, we conducted correlational analysis to investigate whether the video stimuli were assigned similar ratings when presented with and without sound. Results of these analyses are discussed in the light of qualitative inspection of the data.

3.1. The object reference function

As apparent from Figure 2, scores on the Refer-to-Object question range throughout the entire spectrum of certainty ratings. The mean score over all stimuli is 4.06 ($\pm$1.02) for the Sound-On condition and 3.78 ($\pm$1.75) for the Sound-Off condition. From a paired samples t-test, this difference appears significant ($t_{448}=1.80$, $p<0.01$, $d=0.31$). However, note that the effect size is very small: the difference in means comprises less than a third of the standard deviation. Thus, although gestures are more often judged to refer to an object when perceived in the context of speech, their referential function is almost equally easily recognized without access to the verbal channel. This is further corroborated by the correlational analysis. A Pearson test reveals that the ratings in the Sound-On and Sound-Off conditions were strongly correlated $r_{148}=0.65$ ($p<0.01$).

As apparent from Figure 2, there is a difference between the conditions in terms of the variance in the data. The frequency distribution of scores in the Sound-Off condition is much steeper than the one corresponding to the Sound-On condition ($r$ with sound: 1.02, $z$ without sound: 0.75). From this observation, it appears that the presence of speech does not directly contribute to the qualitative interpretation of the gesture, but rather decreases uncertainty among observers. The high rate of “cohesive gestures” [17] in the stimulus set plausibly relates to this finding. Many video fragments contain gestures that refer to an object by referring to a region of space that has previously been associated with a certain landmark object. In the absence of the verbal channel, such gestures may have the appearance of non-representational movements of the hands (e.g. beat gestures; [4]) or, more generally, may not be easily recognized as carrying a referential function.

For some gestures in the stimulus set, the referential status is not clear-cut even in the presence of speech. With regard to trajectory-tracing gestures that co-occur with sentences like over there you should go to the left; there is generally low consensus whether or not the hand refers to an object or person. Though such gestures can be thought of as having a referential component, where the hand ‘embodies’ the interlocutor (cf. [18]), they were often judged to be merely attributive by nature, tracing the movement of some entity without actually referring to it. Thus, for many of our stimuli, responses to the Refer-to-Object question were not entirely homogeneous across participants, regardless of the presence or absence of speech.

3.2. The shape representation function

Mean ratings on the question Depict-Shape do not differ to a statistically significant degree across conditions ($t_{148}=1.8$, $p=0.075$, $d=0.052$). In fact, correlational analysis suggests that gestures were attributed approximately equal ratings on this question when presented with and without sound ($t_{148}=0.84$, $p=0.001$). Figure 3 presents these data graphically. We see that in the Sound-On condition, the distribution of mean scores on the stimuli has two peaks: one close to the ‘certainly not’ pole, and one corresponding to more certain ratings. Thus, for a large portion of the gestures in the stimulus set, participants were rather certain whether or not flatness or size of some object was represented. In the Sound-Off condition, the histogram’s shape more closely approaches a normal distribution. Here we see that when the sound was muted, the overall uncertainty rate with respect to the Depict-Movement question was substantially higher. A likely explanation is that the SaGA corpus, from which the stimuli were extracted, contains many tracing gestures, i.e. gestures where the hand traces a line through the air as it holding a drawing utensil. Such gestures can be semantically ambiguous: tracing handshapes can be used either to draw the shape or outline of some object, or to depict the movement of some object in space [18, 19]. In the context of a verbal utterance such as you will see a curved river, a horizontal tracing gestures will be given a different interpretation than in the context of the utterance you have to turn left and then right like this (static-

![Figure 2: Ratings on the question Refer-to-Object with and without sound](image-url)
attributive in the first case, processual in the second). As reflected by the relatively high degree of gesture stimuli in the middle region of Figure 3a, many tracing gestures therefore remain underspecified in the Sound-Off condition.

3.3. The movement representation function

Scores on the Depict-Movement question diverge to some extent between conditions, but in a different direction than we have seen thus far: the mean rating across all stimuli is higher in the Sound-Off condition than in the Sound-On condition (see Figure 4). This difference is statistically significant ($t_{448} = -15.9$, $p<.001$), but the effect size is small ($d=.28$). Ratings on the stimuli correlate significantly across conditions ($r_{447} = .55$, $p<.001$).

According to these data, not many gestures in the corpus clearly depict the movement of an object or person. Strikingly, direction-tracing gestures such as those discussed earlier (e.g. those co-occurring with go to the left) are not consistently judged as depicting movement. Moreover, likely due to the ambiguity of tracing gestures discussed in the previous section, the Sound-Off condition yielded a high rate of uncertain responses. Many tracing gestures were judged to be potentially depicting some movement when presented in the absence of speech, but were clearly perceived as part of the act of reference to an object (i.e., non-processual in their semantics) when presented with the audio on. Thus, in some cases, the speech is needed to decide whether a tracing gesture has a referential or attributive function. Overall however, we see that the scores in the two conditions correlate strongly, suggesting that this dependence is more of an exception than a rule.

3.4. The meta-communicative function

The three functions discussed so far relate to the domain of semantics: they involve gestures' ability to refer to entities and situations and attribute properties to them. The fourth and final function taken into consideration here has a rather different nature. We here look at gesture’s ability to signal that the speaker experiences difficulty in formulating his utterance, a type of action sometimes referred to as ‘own communication management’ [20]. This section inquires whether access to speech is necessary to understand that a gesture is meta-communicative by character. As clear from Figure 5, the pattern in the data we find here is not substantially different from what we have seen above. Mean scores are differently distributed in the two conditions ($t_{448}=4.2$, $p<.001$), with higher scores in the Sound-On conditions, but the effect size is marginal ($d=.20$). The scores in the Sound-On and Sound-Off

![Figure 3: Ratings on the question Depict-Shape with and without sound](image)

![Figure 4: Ratings on the question Depict-Movement with and without sound](image)
conditions are again strongly correlated ($r_{147}$=53, $p<0.001$).

These data are largely in line with the hypothesis that own communication management is associated with particular formational patterns (finger snapping or rapid cyclic movement of the hands, cf. [20]): the presence or absence of speech may play a role, but is not a crucial factor in deciding whether a given gesture signals own communication management. Nonetheless, the shape of a gesture alone is not always enough to recognize the gesture as having this function. Cyclic gestures, for instance, are not only linked to word search, but also to different contexts, such as encouragement of the interlocutor to continue speaking [21]. In accord with this underspecificity, we see that the right-most region of the distribution in Figure 5A, which represents the most certain/positive responses on the Word-Search question, is virtually empty in the Sound-Off condition (6 stimuli with mean scores above 5), but does have a substantial number of responses in the Sound-On condition (25 stimuli with mean scores above 5). This indicates that even relatively fixed patterns in gesticulation, such as the cyclic gesture and other conventional ways of meta-communicative signaling, are to some degree dependent on verbal context. Nonetheless, a substantial overlap exists in the response profiles across conditions. It appears, as we have seen before, that the judgments in the presence of speech are qualitatively similar, but more assured, than those in the video-only condition. In other words, the influence of the verbal channel on the perception of the gestures is largely limited to the reduction of uncertainty.

4. Discussion and conclusion

Despite ample attention to the question of speech-gesture dependence, it has to date remained relatively unclear how the meaningful qualities that are inherently present in speakers’ gestures are best characterized. Here, we have argued that a functional (linguistic) view on gestural meaning could provide a fruitful level of abstraction to point out which aspects of gestural meaning can be understood without speech. In an internet-based perception study, we have compared people’s perception of gestures with and without access to speech, with respect to four specific functions: object reference, static attribution, dynamic attribution and signaling of own communication management.

Our results suggest that access to the verbal channel does not have strong influence on the perceived functions. All effect sizes were small (between one third and one fifth of a standard deviation) and response profiles show a great deal of overlap across conditions for all functions investigated. Note that the reported $p$-values, of which three out of four were statistically significant, can be misleading. The substantial size of our stimulus set (449 stimuli) yields very high statistical power and an increased risk of false positives. Therefore, in line with current trends in statistics, our conclusions are primarily based on the reported effect sizes and correlations.

A second finding is that the variance in the responses between conditions diverged strongly – it was consistently found that participants were more hesitant and inconsistent in their responses on the video-only stimuli than on those with video and audio. Thus, rather than having a strong influence on the ‘direction’ of the functional interpretation of the gestures, the presence of speech appears to play a role in reducing uncertainty among raters.

These findings raise theoretical as well as methodological considerations. First, they suggest that the question to what degree gestures constitute an independent (sub)system of communication hinges strongly on the level of abstraction one adopts when characterizing their meaning. Using somewhat schematic, functionally oriented questions, we find that the interpretation of gestures is not as dependent on speech as previously supposed on the basis of experiments which ask very concrete questions (e.g., what is the lexical affiliate of this gesture?). This is an important insight, because it informs the level of description appropriate for understanding how speech and gesture intersect: rather than seeking to understand gesture as a system that functions akin to speech (with a full repertoire of very specific form–referent mappings), it is probably better to ask to what extent speech and gesture may have the same kind of communicative role in a given discourse setting, and what formational patterns these roles correlate with in either modality. Thus, the existence of some form of ‘inherent meaning’ of gestures (i.e., speech-independent form–meaning associations) cannot be dispensed with, but should be sought for on a higher level of abstraction than in the case of verbal meanings.

Apart from these theoretical considerations, the results presented here may have methodological implications for the longstanding debate whether annotation practices should be performed with or without audio. Whereas most traditional coding schemes require access to the verbal channel in order to be applicable [3, 22], some contemporary approaches pursue the view that the recognition of meaningful patterns in gestural expression entails at least one round of annotation with the audio turned off ([23, 24]). The high degree of overlap between the Sound-On and Sound-Off conditions in our study supports the latter approach: according to the results reported, it can be fruitful to annotate basic-level functional categories of gestures on the basis of video only. This could benefit the objectivity of the research, because gesture interpretation will be less biased by the co-occurring speech. When dealing with more detailed research questions, however, additional annotation remains necessary. Access to the verbal channel can help the annotator to disambiguate – or specify the meaning of – the gestures of interest. The potential presence of intrinsic high-level meanings in the gestures investigated in this paper, hence, does not suggest that speech and gestures are to be approached as independent systems. Rather, the divergence between our and earlier findings seems
to point out that the further we go along the cline from schematic to lower-level layers of meaning, the more important it is to annotate, and think of, speech and gesture in the context of one another.

5. Acknowledgements

The first author is grateful to support from the Netherlands Scientific Organization (NWO; grant PGW-12-39) and the German Academic Exchange Service (DAAD; scholarship 91526618 - 50015537). Moreover, this research received support by the German Research Foundation (DFG) in the Collaborative Research Center 673 “Alignment in Communication” and the Center of Excellence 277 “Cognitive Interaction Technology” (CITEC).

6. References