Keeping an eye on gestures: Visual perception of gestures in face-to-face communication

MARIANNE GULLBERG AND KENNETH HOLMQVIST

Since listeners usually look at the speaker's face, gestural information has to be absorbed through peripheral visual perception. In the literature, it has been suggested that listeners look at gestures under certain circumstances: 1) when the articulation of the gesture is peripheral; 2) when the speech channel is insufficient for comprehension; and 3) when the speaker him- or herself indicates that the gesture is worthy of attention. The research here reported employs eye tracking techniques to study the perception of gestures in face-to-face interaction. The improved control over the listener's visual channel allows us to test the validity of the above claims. We present preliminary findings substantiating claims 1 and 3, and relate them to theoretical proposals in the literature and to the issue of how visual and cognitive attention are related.

1. Introduction

Gestures occurring during speech — referred to by Kendon (1988) as gesticulation — have been shown to have an information value that can be exploited by speakers and listeners alike. For instance, speakers who are also language learners use gestures as an instrument to elicit lexical help when communicating with native listeners (Gullberg 1998). Listeners also attend to the information expressed gesturally (see Kendon 1994, for a review). In the so-called mismatch experiments (Cassell, McNeill and McCullough 1999;

Pragmatics & Cognition Vol. 7(1), 1999. 35–63

McNeill, Cassell and McCullough 1994), for instance, listeners were asked to retell a story told to them by a narrator whose gestures did not always correspond to the accompanying speech. The results indicate that the information in the gestural channel is retained by the listeners and integrated with information from the spoken channel into what can be thought of as an intermodal cognitive representation of meaning.

In face-to-face interaction, where most gestures are produced, the listener usually looks at the speaker's face (e.g., Argyle and Cook 1976; Kendon 1990). If this norm were followed, gestural information would have to be absorbed through peripheral vision. Studies of gesture perception, especially in face-to-face interaction, have seldom had adequate control over the listener's visual perception, however, but have largely relied on video recordings from the side of the interlocutors. Despite this lack of precision, at least three claims have been made regarding circumstances under which listeners actually look directly at gestures in interaction. The aim of this study is to test these claims by using eye-tracking techniques, allowing precise control over the listener's visual channel.

The three hypotheses tested here belong to different domains. The first hypothesis pertains to the competition between the need to absorb gestural information and the interactional norm that requires listeners to maintain eye contact with speakers. The second concerns the fact that gestures serve as a complementary information source to be exploited when the speech channel is 'noisy'. The third relates to speakers' wish to direct listeners' attention in interaction.

2. Theoretical background and hypotheses

The properties of vision are as important to gesture perception as the properties of the auditory apparatus are to the perception of spoken language. Directing the eye so that the image of the gesture falls on the small, central fovea, is called *fixating* the gesture. The cognitive process responsible for the selection of fixation targets such as gestures is the so-called *pre-attentive process.* This process operates in parallel across the entire visual field to select the next fixation target. It finds the gesture in the periphery and decides that it is of higher value than the face, and so the gesture is fixated. It has been suggested both that the selection process is a simple reaction to

motion or contrast in the periphery (Theeuwes 1993), and that it is guided by high-level cognitive processes such as linguistic priming (de Graef and Spittaels 1997).

We must also distinguish between *foveal* and *peripheral* perception. It is only in the fovea, which spans less than 2 degrees of the visual field, that we can identify finer structures such as letters and the fine articulation of gestures. The possibility of identifying detailed texture decreases rapidly further out in the visual field (Bruce and Green 1990; Latham 1995). For peripheral perception, light and motion recognition is good but textural recognition poor.

When a listener fixates a gesture, the quality of the information retrieved will consequently be higher than if the gesture were to be perceived peripherally. But this added information value must be high enough to outweigh the strength of the norm that dictates eye contact with the speaker. The speaker's face and gestures (and several other objects with other types of value to the listener) can thus be seen as competing for the privilege of being fixated.

In Sign Language, the manual channel is used to transmit the brunt of all information available. Since it is important that the information conveyed be of high quality, we might have expected signs to be fixated in Sign Language conversations. However, signing interlocutors also have strong reasons to fixate the signer's face (Liddell 1980). First, anecdotal reports suggest that the social norm for maintaining eye contact is strong also in signed conversations.[1] Second, important grammatical functions are expressed facially instead of manually in Sign Language. There is in fact experimental data to suggest that peripheral perception is sufficient to enable the signs of Sign Language to be perceived in parallel with overt face fixation (Swisher 1990; Swisher, Christie and Miller 1989). In addition, Siple (1978) has proposed that signs requiring finer articulation are performed closer to the face in order for interlocutors to be able both to maintain eye contact and perceive the signs.

In contrast, in spoken interaction gestures are virtually always accompanied by speech, and speakers therefore do not have to adapt the place of articulation to the perceptive needs of the listener. As a consequence, listeners may be faced with a large proportion of gestures articulated so far out in the periphery that they have to leave the speaker's face and instead fixate the gestures in order to perceive the information expressed. The first

hypothesis to be tested assumes that listeners will fixate the speaker's face unless the gesture is performed in the periphery, in which case listeners will redirect their gaze towards the gesture.

> *First hypothesis*: Peripherally articulated gestures are fixated more often than centrally articulated gestures.

The information value of gestures may also motivate fixations of gestures when the spoken channel is not sufficient for comprehension, as has been suggested by Rogers (1978). This possibility was investigated by Rimé, Boulanger and d'Ydewalle (1988 cited in Rimé and Schiaratura 1991). Subjects were presented with video recordings of an interpreter telling the same story in three different conditions: the listeners' own language (Flemish), in a language partially understood by the listeners (French), and in a language not understood at all (Russian). The subjects' eye movements were recorded, and the results showed that in the native condition, where comprehension was good, relatively few gestures were fixated (a mean of 6.7%). In contrast, in the French and Russian conditions, subjects fixated more gestures (11.2% and 19.6%, respectively). The results indicate that the quality of speech influences the listener's attention to gestures that the 'noisier' the speech channel or the lower the comprehension, the more gestures are fixated. Language learners also produce 'noisy' speech in their attempts to communicate. The second hypothesis was set up to test if native listeners faced with language learners would behave like the subjects in Rimé et al.

> *Second hypothesis*: Gestures are fixated more often when the comprehension of the oral channel is hindered. Native viewers thus fixate proportionally more gestures when listening to non-native narratives than when listening to native narratives.

In studies of interaction, it has been suggested that speakers have various means at their disposal to indicate to a listener that a gesture is communicatively relevant. Verbal pointers or framing devices, as in "then the guy went like this: [gesture]" (Tuite 1993: 94), have been proposed as one method. However, as shown in Streeck (1994), listeners do not necessarily look at gestures framed in this way. Goodwin (1986), Streeck (1993), Tuite (1993) and others have shown that speakers may instead indicate the information value of a gesture by looking at it themselves, by *auto-fixating* it.

When speakers look at their own hands and disrupt eye contact with the listener, the gesture becomes salient. This is a strong indicator that the listener should also direct his or her gaze towards the hands. It has been suggested that auto-fixation always leads to listener-fixations of the same gesture (Streeck 1993; Tuite 1993). These claims have been made without reference to quantitative data, however, and are based on informal side observations of gaze rather than on measures of actual fixations. The third hypothesis has been set up to test these claims under circumstances where the visual channel is under control.

> *Third hypothesis:* Listeners fixate gestures more often than average if these gestures have been auto-fixated by the speaker.

3. Data collection

The set-up consisted of a dyad with a narrator-speaker and a listener-viewer. All subjects received oral instructions. The narrator-speaker was asked to memorize a printed cartoon, and then to retell the story orally to a listener with the stimulus picture removed. The narrator was instructed to ensure that the listener understood the story and the punch-line. No time constraint was imposed on the memorization task. In practice, however, narrators spent only 2–3 minutes looking at the cartoon. Meanwhile, the listener was fitted with an eye-tracker and was informed that s/he was going to be told a story. The importance of understanding the story was stressed, and the listener was encouraged to interact freely with the narrator to achieve this, for instance by asking questions of clarification. As the narrator-speaker retold the story, the listener's visual field and foveal fixation point was recorded on video.

Eight subjects aged between 25 and 50 years participated on a voluntary basis in the study. They were grouped according to their first language so that there were four native speakers of Swedish and four native speakers of French. Two native Swedes, one female and one male, and two native speakers of French, one female and one male, played the role of listener-viewer. Two native Swedes, both female, and two native speakers of French, both female, acted as narrator-speakers. The subjects were not acquainted prior to the experiment.

The four speakers told the story in Swedish to the Swedish listeners, and in French to the French listeners. The speakers were thus required to tell

the story both in their first and in their second language (L1 and L2, respectively). The listeners were told that they would hear both a native and a non-native version of the story.

Two speakers had received no formal instruction in their respective L2s, but had acquired the languages naturalistically in the countries where they are spoken. The third speaker had received only 60 hours of instruction in the L2, but was actively using the language in social contacts. The fourth speaker, finally, had studied the L2 for six years as a foreign language, but had not been using it actively. Despite these varying backgrounds, all subjects were at intermediate proficiency levels.

The design described allowed the subjects to act as their own controls, both in terms of language proficiency, and individual variation in gestural behavior.

The subjects in recordings 1a and 1b were asked to repeat the task six weeks later, which resulted in recordings 1c and 1d. These last two recordings were made to compensate for the lack of sound in 1a and 1b due to equipment failure. They also served the purpose of verifying that the Swedish listener-viewer's gaze behavior was consistent on two different occasions. Table 1 summarizes the recordings.

The eye tracker used is an SMI iView 50 Hz pupil and corneal reflex video imaging system. The eye tracker consists either of a headset, which allows the subject freedom of motion of the head (Figure 1), or of a remote

Table 1. *The ten recordings. L1 = native language; L2 = non-native language. Listeners: 1–4. Speakers: α–δ.*

Recording	*Listener*	*Speaker*	*Language*	*Eye tracker*
1a	Sw Wom 1	Fr Wom α	Swedish L2	Headset
1b	Sw Wom 1	Sw Wom β	Swedish L1	Headset
2a	Fr Man 2	Sw Wom β	French L2	Headset
2b	Fr Man 2	Fr Wom α	French L1	Headset
1c	Sw Wom 1	Fr Wom α	Swedish L2	Remote
1d	Sw Wom 1	Sw Wom β	Swedish L1	Remote
3a	Fr Wom 3	Sw Wom γ	French L2	Headset
3b	Fr Wom 3	Fr Wom δ	French L1	Headset
4a	Sw Man 4	Fr Wom δ	Swedish L2	Headset
4b	Sw Man 4	Sw Wom γ	Swedish L1	Headset

eye tracker placed on a table in front of the subject. The output in the form of video recordings shows the listener's visual field and a moving marker for the listener's foveal attention.

Figure 1. *The SMI iView headset.*

The equipment and its effect on the interaction may potentially compromise the ecological validity of the collected data. A post-test questionnaire was distributed a) to ensure that gesture was not identified as the object of study, and b) to ascertain whether subjects were disturbed by the equipment. The test contained open-ended questions such as *What do you think the point of the study was?*

With regard to the effect on the production of speech and gestures, the questionnaire showed that nobody identified gestures as the object of study. An additional guarantee for the validity of the data is that both speech and gesture production in this study is quantitatively and qualitatively similar to the data in a study with a similar set-up performed without the eye tracker (Gullberg 1998). The distributions of feedback signals, of gesture types, etc., are similar and do not appear to have been affected by the addition of the eye tracker in this study.

With regard to the effect on visual behavior, all subjects, speakers and listeners alike, declared that the equipment did not disturb them. Although all subjects were aware of the objective of the task, i.e. to measure eye movements, this knowledge does not appear to have affected the listeners' visual behavior. The data include fixations of socially unacceptable areas which

suggests that the listeners tended to forget about the apparatus. In sum, the presence of an interlocutor and the pressure of the task seem to have prevailed over the potential awkwardness of the situation, so that subjects obeyed normal interactional conventions despite the experimental context.

4. Coding

4.1. *Gestures*

Gestures are narrowly defined as movements of the hand(s) and/or arm(s) performed spontaneously and unwittingly by the speaker during speech (Kendon 1988; McNeill 1992). This excludes so called emblems, i.e. culture-specific lexicalized gestures such as the V-sign, and self-regulators, i.e., self-grooming gestures such as playing with strands of hair.

Gestures have been coded for their *place of articulation* within gesture space. Central gesture space, as depicted in Figure 2, is where the speaker's gesture production is mainly performed, and it can be delimited by the torso and the length of the lower arms (Haukioja 1992; McNeill 1992). Figure 2 shows the speaker's central gesture space as a white rectangle. Peripheral gesture space is everything outside this rectangle.

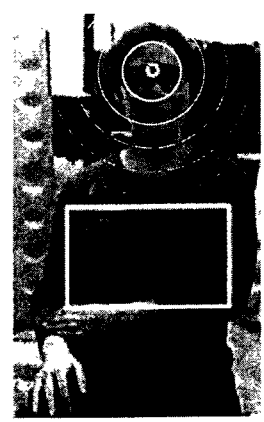


Figure 2. *The speaker's central gesture space (the rectangle), and the listener's fixation (the small circle) and visual periphery (bigger circles).*

Figure 2 represents the canonical case in interaction. The listener's gaze is directed at the speaker's face, as shown by the foveal gaze mark seen as a small white circle. The speaker produces gestures in central gesture space (the rectangle), and the listener has to perceive these gestures through peripheral vision.

Gestures have been coded for place of articulation as Central (C), Horizontal Peripheral (HP), and Vertical Peripheral (VP). Central gesture space has been used as the defining criterion for centrality, with the effect of posture and perspective in the videos taken into consideration. As a consequence, only gestures *well* outside of central gesture space have been coded as peripheral, whereas borderline cases have instead been labeled as central. In cases where a continuous movement occurs both in central and peripheral gesture space, the gesture has been considered as blended and contributed 0.5 central and 0.5 peripheral to frequencies.

All gestures were also coded for *manner of articulation* or *gesture type*, using the coding system proposed by McNeill (1992), as iconics, metaphorics, deictics, and beats.

In addition to this coding, gestures were marked as 'auto' when they were looked at by the speakers themselves.

4.2. *Fixations*

The listener-viewers' fixations were measured from the video recordings. Fixations were identified on the basis both of a temporal and a spatial criterion. In order to count as a fixation, we required the eye marker to remain for at least 80 ms directly *on* the fixated object (hand, arm, face, etc.).

Most studies regarding the lower duration limit of fixations that allow recognition have been made in reading research. The critical break-off point is found in the interval 100–150 ms (Morrison 1984). Neurological studies of fixations and recognition during reading confirm this estimate (Posner, Abdullaev, McCandliss and Sereno, forthcoming). Empirical data on fixation length from a study of eye movements in 3D scenes show that very few fixations, or only 2%, are shorter than 100 ms (Henderson and Hollingworth, forthcoming). Our lower limit of 80 ms thus excludes only that small number of fixations where it is doubtful that the object being fixated has been registered cognitively. In fact, no fixations were shorter than 120 ms in our data. However, fixation time has not been considered as a dependent variable

below, since we are only interested in whether the fixation of the gesture is long enough for recognition to have occurred.

With regard to the spatial criterion, there is little ambiguity as to the object of fixation. The vast majority of fixations rest directly on the speakers' faces. Fixations of gestures are also made directly on the gesture, in most cases the hand. Fixations of non-gestures outside the face were never made in the presence of a gesture in progress. We can therefore rule out the possibility of parafoveal perception of gestures during non-face fixations. In our study, gestures were either fixated or projected onto the peripheral part of the retina during a face fixation.

As a consequence of applying this strict spatial criterion, and combining it with the above recognition criteria for fixation length, we obtain a binomial distribution in the data: each gesture is either fixated or not. Furthermore, we calculate proportions of fixated gestures relative to produced gestures in the various categories, thereby neutralizing the quantitative variations in gesture production. We have employed a test of significance of differences between proportions that is mathematically equivalent to the χ^2–test under one degree of freedom.

5. Data description

5.1. *Gestures*

The data on the speakers' gesture production are shown in Table 2. The four speakers, α–δ, have been indicated using four different shades.

Speakers produce most gestures per minute in their non-native language (proficiency condition L2). The only exception is speaker δ in recording 4a, where the non-native condition results in less gestures per minute than the native condition (recording 3b). This is because the learner hardly speaks at all but stays silent and immobile. According to popular belief, the French subjects would be expected to gesticulate more than the Swedes. The table shows this to be erroneous, however, since the French subjects perform both more *and* fewer gestures than the Swedish subjects. Instead, the proficiency condition influences the amount of gestures produced to a greater extent than cultural membership (cf. Gullberg 1998).

Table 2. *The production data.*

Recording	*Speaker*	*Proficiency condition*	*Tot time of narrative*	*Gestures per minute*	*Peripherally produced gestures*
1a	α French	L2	4:05.80	17.8	8.2%
[redacted]					
2b	α French	L1	1:38.78	11.5	3.4%
1c	α French	L2	2:59.27	13.0	2.6%
[redacted]					
3a	γ Swedish	L2	2:12.57	18.0	16.2%
3b	**δ French**	**L1**	**1:18.16**	**19.2**	**4.0%**
4a	**δ French**	**L2**	**4:25.00**	**5.4**	**8.3%**
4b	γ Swedish	L1	1:02.26	7.7	0.0%
Average			2:49.16	12.1	13.0%

There is also considerable variation in the proportion of peripherally produced gestures, from no peripheral gestures at all to almost every fourth gesture being articulated in the periphery. Peripheral gestures are a measure of the expanse of gestures. Again, popular belief suggests that the French would use more expansive gestures than the Swedes. This does not hold in our data, however. The Swedish subjects use more peripheral gestures than the French subjects in three out of four recordings. Rather than culture, individual propensity determines the use of peripheral gestures.

5.2. *Fixations*

The listeners fixated the speakers' face, gestures and a few other objects.

Table 3 shows that listeners on average spend only 1.6% of the time looking at other things than the speaker's face. This corroborates earlier observations of the strong tendency to maintain eye contact in face-to-face interaction (Argyle and Cook 1976). On average 44% of the time spent outside the face is dedicated to gestures, whereas 56% of the time is spent looking at other things.

Table 3. *Fixation time (in percent of the total duration of the narrative) spent on gestures, on other objects, and outside the speaker's face.*

Recording	*Time on gestures (%)*	*Time on other (%)*	*Σtime outside the face (%)*
1a	1.78%	0.00%	1.78%
1b	0.63%	0.45%	1.08%
2a	1.23%	1.82%	3.05%
2b	1.29%	0.76%	2.05%
1c	0.18%	0.33%	0.51%
1d	0.25%	0.08%	0.33%
3a	0.00%	0.18%	0.18%
3b	0.00%	0.00%	0.00%
4a	0.30%	2.66%	2.96%
4b	0.00%	0.00%	0.00%
Average	0.71%	0.90%	1.61%

Although this paper concerns gesture fixations, a list of the other fixated objects, as shown in Table 4, may help in understanding the characteristics of the distribution of fixations in face-to-face interaction.

A large number of fixations were on empty space; they were labeled 'avoidance fixations'. All but one were consistently directed at the same area, somewhat to the left of the speaker's face, as illustrated in Figure 3.

Table 4. *All fixations of objects other than the speaker's face or gestures.*

Fixated objects	*# of fixations*	*average fixation time*
Empty space (avoidance fixation)	25	240 ms
Breast	17	170 ms
Resting hand, arm, or foot	12	280 ms
Listener's own visual deixis on foot while talking	1	320 ms
Listener follows deictic gesture, but after a long delay	1	240 ms
Speaker's chair	1	120 ms
Speaker's collar	1	120 ms

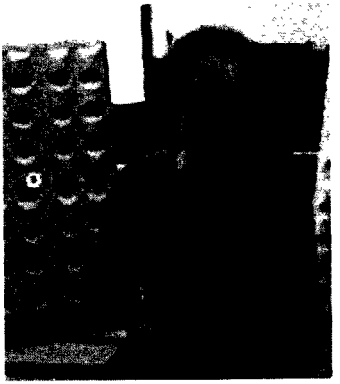


Figure 3. *Avoidance fixation.*

The persistent return to the same area suggests that even avoidance fixations are not random. The majority of the avoidance fixations were made by listener 2, a male subject obviously uncomfortable with sustained eye contact, while a few such fixations were made during speech planning. All breast fixations were made by the male listeners 2 and 4. In fact, the listeners in recordings 2a and 4a were responsible for 85.5% of the fixations of non-gestures.

The individual differences in fixation behavior towards gestures, as shown in Table 5, clearly separate listener 2, the French male listener (highlighted in the table), from the others. Listener 2 on average fixated 27.2% of the gestures produced in recordings 2a and 2b. Listeners 1, 3, and 4, on the other hand, fixated on the average 3.4% of the gestures they encountered. When listeners 1, 3, and 4 are grouped together, a test of homogeneity shows that their behavior is significantly similar ($p \leq .05$). In other words, the probability that the similarity depends on chance alone is equal to or less than five percent. Moreover, when the behavior of the group is compared to that of listener 2, the difference in fixation proportion is significant ($p \leq .001$). Henceforth, all calculations of listener-fixations will include a comparison between the behavior of the group 1, 3, and 4, and that of listener 2.

Some evidence for consistency in listener behavior over time is also seen in Table 5, viz. for the one subject, listener 1, who performed the task twice. Listener 1 showed no significant difference in her gaze behavior between the recordings 1a/1b and 1c/1d, which were made 6 weeks apart.

Table 5. *Proportion of fixated gestures in each of the ten recordings. Listener 2 highlighted.*

Recording	*# of produced gestures*	*# of gestures that the listener fixated*	*Proportion*
1a	73	4	5.4%
1b	36	2	5.6%
2a	58	15	25.9%
2b	19	6	31.6%
1c	39	1	2.6%
1d	20	1	5.0%
3a	40	0	0.0%
3b	25	0	0.0%
4a	24	1	4.1%
4b	8	0	0.0%
Total	341	30	8.8%

6. Results and discussion

6.1. *Central and peripheral gestures*

First hypothesis: Peripherally articulated gestures are fixated more often than centrally articulated gestures.

Table 6 shows the proportion of fixated central and peripheral gestures. Both groups of listeners have a higher proportion of gesture fixation for peripherally articulated gestures, but we find no significant difference between centrally and peripherally articulated gestures for the French listener 2. For the group 1, 3, and 4, the difference in fixation proportions between central and peripheral gestures is significant ($p \leq .001$). The difference remains significant when we examine all listeners together ($p \leq .01$). The conclusion is that while listeners 1, 3, and 4 fixated more of the peripheral gestures, listener 2 may have fixated central and peripheral gestures to the same extent, with only a weak tendency towards favoring peripheral gestures. Listener 2's behavior may in part be explained by the fact that he also displays a large amount of avoidance fixations. It is possible that a number

Table 6. *Proportion of fixated central and peripheral gestures.*

Listener(s)	*Region*	*# of gestures*	*# of these gestures that the listener fixated*	*Proportion*
2	Central	61	16	26.2%
	Peripheral	16	5	31.2%
1, 3, and 4	Central	235.5	5	2.1%
	Peripheral	28.5	4	14.0%
All	Central	296.5	21	7.1%
	Peripheral	44.5	9	20.2%

of his fixations of gestures are part of his tendency to avoid looking at the speaker's face. Since his avoidance fixations can be assumed to be evenly distributed over all produced gestures, most of which are centrally articulated, this may explain why listener 2 shows an over-representation of fixations of centrally performed gestures.

When the two axes of the peripherally articulated gestures are considered, as in Table 7, the data for listener 2 are not sufficient for statistical significance. For the group 1, 3, and 4, the difference in fixation proportions between horizontal and vertical gestures is significant ($p \leq .01$). When all listeners are considered together, we find that peripheral gestures performed in the horizontal dimension are fixated significantly more often than centrally articulated gestures ($p \leq .05$). Peripheral gestures performed in the vertical

Table 7. *Proportions of fixated peripheral gestures in the two scalar dimensions.*

Listener(s)	*Peripheral dimension*	*# of gestures*	*# of these gestures that the listener fixated*	*Proportion*
2	Horizontal	13	3	23.1%
	Vertical	3	2	66.7%
1, 3, and 4	Horizontal	19	0	0.0%
	Vertical	9.5	4	42.1%
All	Horizontal	32	3	9.4%
	Vertical	12.5	6	48.0%

Table 8. *Summary of comparisons between central and peripheral gestures in the vertical and horizontal dimensions.*

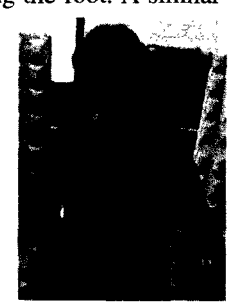
Compared dimensions	*p-value*
central vs. horizontal (all subjects)	$p \leq .05$
vertical vs. horizontal (all subjects)	$p \leq .01$
central vs. vertical (all subjects)	$p \leq .001$

dimension are also fixated significantly more often than central gestures ($p \leq .001$), and also more often than horizontally articulated peripheral gestures ($p \leq .01$). These comparisons are summarized in Table 8.

Thus, for the ensemble of the data, a hierarchy of fixation behavior can be detected with vertical peripheral gestures being fixated more often than horizontal peripheral gestures, which in turn are fixated more often than centrally articulated gestures:

PERIPHERAL VERTICAL > PERIPHERAL HORIZONTAL > CENTRAL

A tempting explanation for the tendency for vertical peripheral gestures to be fixated lies in the fact that the human visual field has greater horizontal width than vertical span. This leaves greater possibilities for peripheral gestures performed on the horizontal axis to be perceived by peripheral vision. A gesture performed in the lower vertical periphery, however, risks being missed unless fixated by foveal vision. In the data at hand, all gesturesinvolving the foot of the speaker consistently attract the listeners' fixations, as exemplified in Figure 4. The speaker is retelling a scene in the narrative where a doctor is writing a prescription by foot, and the speaker is indicating the foot. A similar finding is presented by Streeck (1994). On the



et / eeh / **mis** le [le ⟨t⟩]
[le] mm ***pen***

and / eeh / **put** the [the ⟨t⟩]
[the] mm ***pen***

Figure 4. *A vertical gesture fixated by the listener.*

one hand, a large, two-handed horizontal peripheral gesture is *not* fixated by the listener although it is accompanied by an oral "pointer" or a deictic expression of the type "about this size" (p. 243). The information conveyed is peripherally perceived by the listener. On the other hand, a gesture performed "just above [the speaker's] lap" (p. 244) in connection with mention of a skirt, receives the listener's direct visual attention. Streeck's explanation for why the latter is fixated is that rapid movements attract gaze. Aside from the fact that there is no way of determining whether we are dealing with real fixations since the analysis is based on the informal observation of the general direction of the interlocutor's gaze, movement per se appears to be an unlikely candidate as a foveal attractor. All gestures involve movement in the visual field, but not all of them are fixated. Moreover, determining what 'rapid' means is very difficult. It seems more plausible that the lap gesture is fixated because it is articulated far out in the listener's peripheral visual field, and requires a re-orientation of gaze in order to be perceived at all.

Another reason why the vertical peripheral gestures in the data are more likely to be fixated may be that they are all examples of *concrete deictic* gestures. Such pointing gestures have as their built-in function to direct listeners' attention to the target of the gesture. This deictic function has been referred to as *demonstratio ad oculos* (Bühler 1934) or *indicatio ad oculos* (Slama-Cazacu 1976). The normal expectation, therefore, is that listeners look at the target of the deictic gesture. In those cases where the target and the gesture itself converge, i.e. when the speaker touches the target, a target fixation will equal a gesture fixation, which is what is exemplified in Figure 4.

The tendency for vertical peripheral gestures to be fixated in our data may also be a task-based effect, however. The foot writing is part of the punch-line of the narrative. The combination of the vertical peripheral position, the concrete deixis, and the punch-line, are likely to serve as a strong cluster of attractors of visual fixation for the listener. At this point, we cannot exclude the possibility that a horizontal peripheral gesture combined with the narrative punch-line would not result in the same number of fixations.

6.2. *Native and non-native speakers*

> *Second hypothesis*: Listeners looking at non-native speakers fixate more gestures than listeners looking at native speakers.

Table 9. *Proportion of fixated gestures in the native (L1) and non-native (L2) conditions.*

Listener(s)	*Proficiency condition*	*# of gestures*	*# of these gestures that the listener fixated*	*Proportion*
2	L1	19	6	25.9%
	L2	58	15	31.6%
1, 3, and 4	L1	89	3	3.4%
	L2	176	6	3.4%
All	L1	107	9	8.4%
	L2	234	21	9.0%

In Table 9, the proportions of fixated gestures have been divided into native (L1) and non-native (L2) speaker production. Contrary to expectations, native listeners do *not* fixate more gestures when they listen to non-native speakers than they do when listening to native narratives (9.0% in L2 vs. 8.4% in L1). There is no significant difference between the conditions and the second hypothesis has to be rejected. Instead, the similarity in behavior in the two conditions is significant, meaning that the likelihood that the similarity depends on chance alone is equal to or less than five percent ($p \leq .05$). This holds also when we look at listener 2 and at the group 1, 3, and 4, respectively. These findings are in contradiction with the results presented by Rimé et al. (1988).

The difference between our results and those presented by Rimé and his colleagues can be accounted for in a number of ways. First, the higher rate of gesture fixations in their study may be a reflection of the fact that they used video-recorded stimuli. Their design did not require the listeners to take interactional norms for eye contact into consideration, leaving them freedom to fixate gestures at their leisure. In contrast, our listeners were faced with real people and had to obey the norms for social interaction and maintain their visual focus on the speaker's face. In view of this methodological difference, it was predictable that our study would result in lower numbers of fixations.

However, the use of video stimuli in Rimé et al. may also have affected the results in other ways. The quality of the stimuli may be different when video recorded than when performed face-to-face. It is not clear how addressing a camera rather than an interlocutor might have affected the

speakers' gesticulatory behavior. At the very least, we can assume that speakers gesticulated less than they would have in real interaction, since we know that the social situation affects both the number and the type of gestures produced (e.g. Aboudan and Beattie 1996; Bavelas, Chovil, Lawrie and Wade 1992). On the other hand, if speakers were told that listeners would potentially be language learners, this might have affected the speakers' use of gesture in the opposite direction, resulting in more gestures. Nothing is in fact known about how native speakers gesticulate when addressing non-native listeners, but in view of the evidence for modified interaction in the oral channel (e.g., Long 1983), it seems fair to assume that some modification occurs.

Second, the lack of difference between the proficiency conditions in our data may be due to the fact that the speech channel was not 'noisy enough' in the L2 condition. In fact, in our data the proficiency level of the language learners was such that serious communicative problems in the oral channel arose only in one case, and this resulted in long silences on the part of the learner (recording 4a). As a consequence, the learners did not gesticulate very much, and in general the listeners could largely rely on the oral channel for information. Moreover, the proficiency level of the learners was only informally defined. Further studies should include subjects whose proficiency level has been determined with formal criteria (such as test batteries and/or evaluation by native judges). Furthermore, subjects should preferably be chosen from low and intermediate levels of proficiency in order to enable reliable assessment of native listeners' gaze behavior towards gesticulating language learners. In addition, Rimé et al. tested non-native listeners faced with languages partially or not understood, whereas our design tested *native* listeners faced with non-native speakers, i.e. an 'accented' variety of their own language. Comprehension of a variety of the mother tongue is likely to be easier than that of a variety of a totally different language.

In view of these observations, the present results on the role of language proficiency may be considered to be inconclusive.

6.3. *Auto-fixation*

Third hypothesis: Listeners fixate gestures more often than average if these gestures have been auto-fixated by the speaker.

Table 10. *Proportion of auto-fixated gestures that were followed by listener-fixation of the same gesture.*

Recording	*# of auto-fixated gestures*	*# of these gestures that the listener fixated*	*Proportion*
1a	2	2	100%
1b	5	2	40%
2a	3	2	67%
2b	3	1	33%
1c	0	0	–
1d	0	0	–
3a	0	0	–
3b	0	0	–
4a	0	0	–
4b	0	0	–
Total	13	7	53.8%

The proportion of auto-fixated gestures that were followed by at least one listener-fixation of the same gesture is shown in Table 10. The statement that auto-fixation leads to listener-fixation (Streeck 1993, 1994; Tuite 1993) is supported in our material. However, as is clear from the descriptive data, not *all* auto-fixated gestures are followed by a listener-fixation. They are, however, significantly more likely to be listener-fixated than other gestures. The difference in proportions is significant ($p \leq .001$) even with our small data set. No calculation has been performed comparing listener 2 and the group 1, 3, and 4 in this instance, as the data set is too small for a meaningful comparison to be made.

It has been suggested in the literature that there is a relationship between iconicity and auto-fixation, as well as between iconicity and *listener-fixation*, but this relationship appears to be little understood. Starting with iconicity and auto-fixation, Streeck (1993) claims that all auto-fixated gestures are iconic. This is an attenuated form of the proposal in Streeck and Knapp (1992), where it is stated that all iconic gestures are auto-fixated. The latter claim is not empirically supported. Only 10.5% of the iconic gestures in our data were auto-fixated. However, our data *do* support Streeck's (1993) claim that all auto-fixated gestures are iconic. Streeck does not offer any explanation for why this should be so, however.

Figure 5. *A speaker's auto-fixation followed by a listener-fixation of the same gesture.*

The auto-fixated iconics in our data are all highly mimetic, in the sense that the speaker has assumed the role of the character in the story and acts out its part. An example can be seen in Figure 5. The speaker is looking at an imagined medical prescription referred to in the narrative. The eye tracking marker shows the listener fixating the speaker's hand holding the prescription, or possibly, the hand being the prescription. The speaker obviously acts as the sales person in the narrative. A mimetic gesture can be technically defined as an iconic gesture in which not only the hands, but the entire body — and specifically the head — are used as articulators. A scale of increasing mimesis can be established on the basis of how many articulators are involved in the performance of the gesture (Gullberg 1998). As can be seen in Figure 5, when the speaker's head becomes an articulator, she looks at her gestures as part of a larger mimetic act. We would like to suggest that it is not iconicity *per se* which determines auto-fixation, but instead the mimetic act which consists of a (mimetic) iconic gesture with the possible inclusion of other articulators, viz. the head and eyes. Mimetic auto-fixation is a reflection of the narrative effort as such, equivalent to quotations or direct speech (Clark and Gerrig 1990).

The common explanation for why auto-fixations are followed by listener-fixations is that auto-fixation is a form of intentional visual deixis by which the speaker indicates the gesture's relevance to the listener. If auto-fixation is instead seen as part of the narrative effort, then an alternative explanation can be given for *listener*-fixations of auto-fixated gestures.

Through the speaker's auto-fixation and mimesis, the listener may be drawn into the mimetic situation, i.e. shifting narrative level, and becoming part of the mimetic act. For instance, the listener in Figure 5 might feel that she is the customer to the speaker's sales person. As such, she acts as she would in the real situation, looking at whatever the speaker is interested in. The possibility that the listener interprets auto-fixation as intentional deixis cannot be categorically ruled out, but it seems just as likely that the narrative context plays an important part in the listener's behavior.

6.4. *Summary of the empirical results*

The empirical findings of this study show that, despite individual quantitative differences in fixation behavior, two conditions have been identified during which the average listener abandons the speaker's face in order to fixate gestures instead. First, gestures performed in the vertical periphery are more often fixated than centrally performed gestures, especially if they are also concrete deictics. This can be assumed to be a function both of the physiology of the human visual field, and of the communicative function of pointing. Second, if speakers fixate their own gestures, listeners tend to follow suit, either as a reaction to visual deixis, or to the switch in narrative level achieved through mimesis.

7. General discussion

The results from this study confirm that it is possible to use eye tracking techniques to test hypotheses about listeners' visual behavior towards gestures without unduly compromising the ecological validity of the results. The main advantage of using this technique is of course the precise control that can be achieved over the listener's fixations. Analyzing fixation behavior from side video recordings, as done in most other studies, is a much less accurate method. The low precision attained makes it impossible to distinguish fixations of gestures from fixations of other objects. Also, many eye blinks may be mistaken for fixations in the lower vertical space, since both cause the eyelid to close.

The eye tracker helps to provide answers to the questions of which and how many gestures are fixated, and it permits us to have some control over

the distinction between foveal and peripheral visual perception. But how can this control further our understanding of what it means to attend to a gesture or to incorporate the gestural information into the intermodal cognitive representation assumed by the mismatch studies?

While fixations are overt physiological events, attention is a cognitive phenomenon. In the following, *attending to* a gesture will be taken to mean the act of directing your focus of consciousness towards the gesture (in the sense of Chafe 1994). Three cases relate fixation to visual attention:

> A1) The gesture is fixated. In tasks that demand much information to be retrieved by visual perception, such as face-to-face communication, the fixated object is with few exceptions also attended to. The possible exception in our data may be those extra fixations listener 2 made of centrally articulated gestures, if he made them, not because he was very interested in the gesture, but because he wanted to avoid the face. Listener 2's attention may at the time of the gesture fixation have been focused on some idea connected to the avoidance of the face, rather than on the gesture.
>
> A2) Attention is on an object, such as a gesture, in the visual field. This object will usually be fixated (the pre-attentive process causes a fixation to be made of the object attended to). In some cases, however, the attended and the fixated object may not be one and the same. Attention is then said to be *covert* (Johnson 1995; Posner, Snyder and Davidson 1980). We cannot rule out that our listeners attended to some gestures without fixating them.
>
> A3) When a listener fixates and attends to the same object (such as the face), then the peripherally perceived objects (such as gestures) are not attended to.

These three cases indicate that fixation and (visual) attention are usually the same — with only a few exceptions. Fixation data therefore provide a measure, albeit partial, of the attention given to gestures.

The expression *attending to gestures* has also been used for the process of including the gestural information in an intermodal cognitive representation of the narrative. The mismatch experiments (Cassell et al. 1999; McNeill et al. 1994), briefly described in the introduction, aimed to show that gestural information is attended to in this sense along with oral material. In these studies, listeners were presented with a video-recording of a person telling a

story. The accompanying gestures had been manipulated so that they were sometimes matched to the verbal content, and sometimes mismatched. A matched gesture would express the same information as that provided in the oral channel. A mismatched gesture, on the other hand, would express contradictory information. The listeners were subsequently asked to retell the story they had just seen. The results show that when there was information in the gesture that was not conveyed orally, e.g. on the manner of movement, then this information was retained by the listeners and retold. Conversely, information expressed orally in the stimulus was sometimes retold gesturally. Finally, when there was discrepancy between the information in the oral and the gestural channels, subjects tried to reconcile the conflicting information either in one or both modalities.

The mismatch experiments provide a partial measure of the number of gestures that are attended to, i.e. integrated into the cognitive representation. It is partial because it is likely that a number of gestures were included in the listeners' representations of the narrative but never activated in the retellings. In other words, more gestures may have been integrated into the cognitive representation than turned up in the retellings.

At the same time, however, there is reason to suspect that the number of gestures integrated into the cognitive representation may have been greater in the mismatch experiment than it would have been in normal interaction. The results show a high proportion of gestures re-emerging in retellings (43% of the gestures expressing manner, 50% of the gestures where the origo or point of departure was manipulated, and 32% of the gestures expressing anaphor, respectively; Cassell et al. 1999). These figures can be compared to the mere 8.8% fixated gestures in our data (in fact, 3.4% for the group and 27.2% for listener 2). The gestural stimuli in the mismatch studies were manipulated, performed consciously by the speaker, and sometimes also mismatched. Their relative unnaturalness is likely to have made them more attractive to the pre-attentive process than naturally occurring gestures. It is therefore possible that the mismatch-listeners fixated, attended to, and integrated into their cognitive representation a considerably higher number of gestures than our listeners. Furthermore, the mismatch stimuli were presented as video recordings. As in Rimé et. al. (1988), the video presentation of stimuli may cause a further increase in fixation frequency as the subjects are not obliged to abide by the norm of looking into the speaker's face.

Fixation data might help elucidate the complex relationships surrounding the mismatch figures. In order to do so, however, the relationship between fixation-attention in the sense of cases A1–A3, and the cognitive representation, has first to be ascertained. There are four logical possibilities that are increasingly likely to be integrated into the cognitive representation:

CR1) The gesture is peripherally perceived and not attended to. We cannot rule out the possibility that some of these gestures are nevertheless integrated into the cognitive representation. The information simultaneously expressed in the oral context provides support that may be suffficient for the integration of peripherally perceived unattended gestures.

CR2) The gesture is fixated but not attended to. Since the gesture is fixated, the gestural information will be of a higher quality than for gestures that are peripherally perceived. The higher quality of the information should not reduce the likelihood that the gesture is included in the cognitive representation, but rather the opposite.

CR3) The gesture is covertly attended to and not fixated. Since the gesture is attended to, it should be more likely that it finds its way into the cognitive representation than those gestures that are not attended to.

CR4) The gesture is both fixated and attended to. It receives a comparatively salient position in the cognitive representation. The overwhelming majority of the fixated gestures in our data ought to fall into this category.

The fixation frequency for the group, 3.4%, is a reliable conservative lower measure of the actual integration of gestures in the cognitive representations of these listeners. As for listener 2, we cannot confidently argue that his 27.2% is also a lower limit of the number of integrated gestures, in view of his high number of fixations of central gestures. Many of these may have been unattended (see A1 above), but may nevertheless not necessarily have been excluded from the cognitive representation (CR2).

Considering the individual differences in fixation frequency, it is reasonable to expect similar differences with respect to the integration of information into the cognitive representation. The roughly 40% reported by Cassell et al. is an average over all subjects, however. In addition, it may be both too high and too low a measure of the number of integrated gestures. We can hypothesize, however, that the information from those gestures that are fixated should be highly over-represented in the mismatch output.

Such a hypothesis could be tested if the mismatch experiments were to be complemented by the use of an eye tracker. In addition, there are at least four methodological advantages to combining eye tracking with the mismatch design. We would know for those gestures that re-emerge in the mismatch output precisely where on the retina they have originally been projected — not only whether they are foveal or peripheral. This would give us an improved opportunity to study what happens to gestures that are peripherally perceived. Also, we could eliminate the risk that potential avoidance fixations of gestures do not convey information to the cognitive representation, if a substantial number of these gestures re-emerge in the mismatch output. For the mismatch studies it would be an advantage to control the visual input. Such control would improve the validity of the constructed stimuli since it would be possible to determine whether the manipulated gestures are fixated more often than the natural gestures. On the same grounds, it would allow control of the effect of the video presentation. If the face is less fixated than in face-to-face interaction, the effect of the video stimulus would be confirmed.

Lunds Universitet

Acknowledgments

We would like to thank the Magnus Bergwall Foundation for support of this research. We also acknowledge the help of our informants, and the useful comments of two anonymous referees on an earlier draft of this paper.

Note

1. As suggested by the editor, lip reading may also condition fixations of the face. This is probably true for conversations between a hearing and a signing interlocutor, but the role of lip reading in conversations between native speakers of Sign Language is unclear.

References

Aboudan, R. and Beattie, G. 1996. "Cross-cultural *similarities* in gestures. The deep relationship between gestures and speech which transcends language barriers". *Semiotica* 111(3/4): 269–294.

Argyle, M. and Cook, M. 1976. *Gaze and Mutual Gaze.* Cambridge: Cambridge University Press.

Bavelas, J. B., Chovil, N., Lawrie, D. A., and Wade, A. 1992. "Interactive gestures". *Discourse Processes* 15(4): 469–489.

Bretécher, C. 1985. *Docteur Ventouse Bobologue.* Paris: Bretécher.

Bruce, V. and Green, P. R. 1990. *Visual Perception: Physiology, Psychology and Ecology.* 2nd ed. Hove, UK: Lawrence Erlbaum.

Bühler, K. 1934. *Sprachtheorie.* Jena: Fischer.

Cassell, J., McNeill, D., and McCullough, K.-E. 1999. "Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information". *Pragmatics & Cognition* 7(1): 1-33.

Chafe, W. 1994. *Discourse, Consciousness, and Time.* Chicago: University of Chicago Press.

Clark, H. H. and Gerrig, R. J. 1990. "Quotations as demonstrations". *Language* 66(4): 764–805.

Goodwin, C. 1986. "Gestures as a resource for the organization of mutual orientation". *Semiotica* 62(1/2): 29–49.

de Graef, P. and Spittaels, O. 1997. "Parallel object processing in real-world scenes". Paper presented at the 9th European Conference on Eye Movements, Ulm.

Gullberg, M. 1998. *Gesture as a Communication Strategy in Second Language Discourse.* Lund: Lund University Press.

Haukioja, T. 1992. "Pointing in sign language and gesture: An alternative interpretation". *Language and Communication* 13(1): 19–25.

Henderson, J. M. and Hollingworth, A. Forthcoming. "Eye movements during scene viewing: An overview". In G. W. Underwood (ed), *Eye Guidance While Reading and While Watching Dynamic Scenes.* Amsterdam: Elsevier.

Johnson, M. H. 1995. "The development of visual attention: a cognitive neuroscience perspective". In M. S. Gazzaniga (ed), *The Cognitive Neurosciences.* Cambridge, MA: The MIT Press, 735–747.

Kendon, A. 1988. "How gestures can become like words". In F. Poyatos (ed.), *Crosscultural Perspectives in Nonverbal Communication.* Toronto: Hogrefe, 131–141.

Kendon, A. 1990. *Conducting Interaction.* Cambridge: Cambridge University Press.

Kendon, A. 1994. "Do gestures communicate?: A review". *Research on Language and Social Interaction* 27(3): 175–200.

Latham, K. J. C. 1995. *Psychophysical Investigations of Human Peripheral Vision.* Ph.D. Diss., The University of Aston in Birmingham.

Liddell, S. K. 1980. *American Sign Language Syntax.* The Hague: Mouton.

Long, M. H. 1983. "Native speaker/non-native speaker conversation and the negotiation of meaning". *Applied Linguistics* 4: 126–141.

McNeill, D. 1992. *Hand and Mind. What the Hands Reveal about Thought.* Chicago: Chicago University Press.

McNeill, D., Cassell, J. and McCullough, K.-E. 1994. "Communicative effects of speech mismatched gestures". *Research on Language and Social Interaction* 27(3): 223–237.

McNeill, D., Levy, E. T. and Cassell, J. 1993. "Abstract deixis". *Semiotica* 95(1/2): 5–19.

Morrison, R. E. 1984. "Manipulation of stimulus onset delay in reading: Evidence for parallel programming of saccades". *Journal of Experimental Psychology: Human Perception and Performance* 10(5): 667–682.

Posner, M. I., Snyder, C. R. R. and Davidson, B. J. 1980. "Attention and the detection of signals". *Journal of Experimental Psychology: General* 109(2): 160–174.

Posner, M. I., Abdullaev, Y. G., McCandliss, B. D., and Sereno, S. C. Forthcoming. "Anatomy, circuitry, and plasticity of word reading". In J. Everatt (ed), *Visual and Attentional Processes in Reading and Dyslexia.* London: Routledge.

Rimé, B., Boulanger, B. and d'Ydewalle, G. 1988. "Visual attention to the communicator's nonverbal behavior as a function of the intelligibility of the message". Paper presented at the Symposium on TV Behavior, 24th International Congress of Psychology, Sydney, Australia.

Rimé, B. and Schiaratura, L. 1991. "Gesture and speech". In R. S. Feldman and B. Rimé (eds), *Fundamentals of Nonverbal Behavior.* Cambridge: Cambridge University Press, 239–281.

Rogers, W. T. 1978. "The contribution of kinesic illustrators toward the comprehension of verbal behavior within utterances". *Human Communication Research* 5(1): 54–62.

Siple, P. A. 1978. "Visual constraints for sign language communication". *Sign Language Studies* 19: 95–110.

Slama-Cazacu, T. 1976. "Non-verbal components in message sequence: 'Mixed syntax'". In W. C. McCormack and S. A. Wurm (eds), *Language and Man: Anthropological Issues.* The Hague: Mouton, 217–227.

Streeck, J. and Knapp, M. L. 1992. "The interaction of visual and verbal features in human communication". In F. Poyatos (ed.), *Advances in Nonverbal Communication.* Amsterdam: Benjamins, 3–23.

Streeck, J. 1993. "Gesture as communication I: Its coordination with gaze and speech". *Communication Monographs* 60(4): 275–299.

Streeck, J. 1994. "Gesture as communication II: The audience as co-author". *Research on Language and Social Interaction* 27(3): 239–267.

Swisher, M. V. 1990. "Developmental effects on the reception of signs in peripheral vision". *Sign Language Studies* 66: 45–60.

Swisher, M. V., Christie, K. and Miller, S. 1989. "The reception of signs in peripheral vision". *Sign Language Studies* 63: 99–125.

Theeuwes, J. 1993. "Visual selective attention: A theoretical analysis". *Acta Psychologica* 83: 93–154.

Tuite, K. 1993. "The production of gesture". *Semiotica* 93(1/2): 83–105.