# Postcards from the mind

## The relationship between speech, imagistic gesture, and thought

Jan Peter de Ruiter
Max Planck Institute for Psycholinguistics

In this paper, I compare three different assumptions about the relationship between speech, thought and gesture. These assumptions have profound consequences for theories about the representations and processing involved in gesture and speech production. I associate these assumptions with three simplified processing architectures. In the *Window Architecture*, gesture provides us with a 'window into the mind'. In the *Language Architecture*, properties of language have an influence on gesture. In the *Postcard Architecture*, gesture and speech are planned by a single process to become one multimodal message. The popular Window Architecture is based on the assumption that gestures come, as it were, straight out of the mind. I argue that during the creation of overt imagistic gestures, many processes, especially those related to (a) recipient design, and (b) effects of language structure, cause an observable gesture to be very different from the original thought that it expresses. The Language Architecture and the Postcard Architecture differ from the Window Architecture in that they both incorporate a central component which plans gesture and speech together, however they differ from each other in the way they align gesture and speech. The Postcard Architecture assumes that the process creating a multimodal message involving both gesture and speech has access to the concepts that are available in speech, while the Language Architecture relies on interprocess communication to resolve potential conflicts between the content of gesture and speech.

**Keywords:** iconic gesture, window into the mind, cognitive architecture, representational gestures, gesture and speech

## Introduction

Human face-to-face communication usually involves not only the exchange of fragments of speech, but also of signals in several other sensory and functional

*Gesture* 7:1 (2007), **21–38**.
ISSN 1568–1475 / E-ISSN 1569–9773 © John Benjamins Publishing Company

communicative modalities,[1] such as facial expression, eye-gaze and gesture. Since the ground breaking work of Kendon (1972, 1980, 2004) and McNeill (1992, 2000), the complex, spontaneous and meaningful hand motions that accompany speech have received a growing attention from semioticians, linguists, and psycholinguists. As many gesture researchers have repeatedly demonstrated, gesture and speech are orchestrated together to form coherent multimodal messages. In the words of Kendon (1972, p. 205) it is "as if the speech production process is manifested in two forms of activity simultaneously: in the vocal organs and also in bodily movement".

An intriguing class of gestures is what McNeill (1992) calls *imagistic* gestures. These are special because they are not conventionalized. While other gestures, such as *emblems* or *quotable gestures* (Kendon, 1990) and *pointing gestures* have conventionalized form-meaning mappings shared within a given linguistic community (Kendon, 1988; Kendon & Versante, 2003; Wilkins, 2003), imagistic gestures have a form-meaning relationship that is 'idiosyncratic' (McNeill, 1992).[2] This implies that the meaning of these imagistic gestures can only be inferred from their spatio-temporal characteristics and the information available in the accompanying speech. For the remainder of this paper, I follow the taxonomy by McNeill (1992), and use the word 'gesture' to refer to imagistic gestures, including both iconic and metaphoric gestures.

Because of the spontaneous, non-conventional form-meaning relationship assumed to be characteristic of such gestures, they are sometimes supposed to provide researchers with a 'window into the mind' (Beattie, 2003; Goldin-Meadow, Alibali & Church, 1993; McNeill, 1992; McNeill & Duncan, 2000), revealing aspects of a thought that the speaker did not necessarily want to share with interlocutors. Other researchers (Enfield, 2005; Kendon, 1972, 1980, 1994, 2004; De Ruiter, 2000, 2003) propose that gestures are *designed* to communicate specific ideas to the interlocutor (together with speech). A third type of relationship between thought and gesture has been put forth by Kita & Özyürek (2003). They present cross-linguistic evidence supporting the idea that gestures can be shaped by the structure of the language being spoken. They have formulated an 'Interface Hypothesis' to accommodate these findings. In their Interface Hypothesis, gesture is the product of a constraint satisfaction or negotiation process between the requirements of representing the underlying imagery as accurately as possible on the one hand, and having the gesture correspond to the verbal concepts being expressed in speech on the other.

## Architectures of the relationship between speech, thought and gesture

These three positions on the relationship between gesture and thought imply three general processing architectures. In this paper, I distinguish between *architectures* and *models*. The architectures and their corresponding diagrams are didactic devices, intended to visualize the core assumptions embedded in certain theories and/or models of gesture. The architectures presented below are not models. When I write 'model' in this paper, I refer to a specific model of gesture processing, and when I write 'architecture', I refer to a class of models or theories that share the same core assumptions about the relationship between speech, thought and gesture.

The architecture implied by the idea that gesture is a window into the mind is called the *Window Architecture*, illustrated in Figure 1. The box labeled 'Formulator' represents all the processing involved in transforming a 'preverbal message' into overt speech (see Levelt, 1989, for a comprehensive overview of these processes).

Here it is assumed that gestures come straight out of the mind. Beattie (2003) writes: "Indeed, these movements of the hands and arms reflect our thinking, like language itself but in a completely different manner. I will argue that such behaviors provide us with a glimpse of our hidden unarticulated thoughts. Movements of the hands and arms act as a window on the human mind; they make thought visible." Although in McNeill's 'growth point theory' (McNeill, 1992), gesture and speech are more intricately linked with each other than the quote by Beattie suggests, McNeill also claims that imagistic information is often expressed 'unwittingly' in gesture: "Gestures exhibit images that cannot always be expressed in speech, as well as images the speaker thinks are concealed" (McNeill, 1992, p. 11). In a discussion about using gesture to do 'mind reading', McNeill writes: "Mind reading is also possible in non-narrative discourse. In the example of this section it is a conversation in

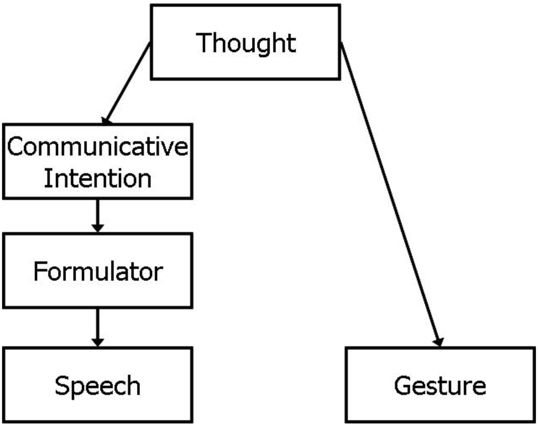

**Figure 1.** The Window Architecture

which a speaker unwittingly reveals something that he was attempting to conceal" (p. 113). Later in that same section, McNeill concludes that "through the gesture we can read his mind to find a meaning which, at this moment, he was attempting to conceal" (p. 116). So according to this view, gestures sometimes act like Freudian slips of the hand, revealing information about a speaker's thought that they did not necessarily want to communicate to their interlocutor.

A different argument by McNeill & Duncan (2000) for gesture acting like a window into the mind is that the process of transforming a thought into a gesture is less complex than the process of transforming a thought into speech. Gestures do not display the same hierarchical (syntactic) complexity that speech does. Gestures often do come in sequences, but these sequences are not hierarchically structured.[3] Also, because gestures are not conventionalized, they need not conform to lexical and morphological conventions, like words do. Because the transformation from thought to gesture is less complex than from thought to speech, gestures "open a 'window' onto thinking that is otherwise curtained. Such a gesture displays mental content, and does so instantaneously, in real time…" (McNeill & Duncan, 2000, p. 143).

The *Language Architecture* implies that the language you speak affects your gesture. Kita & Özyürek (2003) found that the way languages encode information can have consequences for the shape of the gestures produced by speakers of that language. For example, in English, manner (e.g., 'rolling') and path (e.g., 'down') can be expressed together, by using the expression 'rolling down'. Because the path and manner are encoded verbally in the same clause, the authors argue, the gesture that often accompanies the speech 'he rolls down' also contains both manner and path (see the example below for more details). Speakers of Turkish and Japanese, languages in which one has to express manner and path in separate clauses (e.g., 'move down, in a rolling fashion'), also produce separate *gestures* for manner and path. Another example in Kita & Özyürek concerns the description of a cartoon scene in which a cat uses a rope to 'swing across' to the other side of the street. English speakers, who would normally use the verb 'swing' to describe the action of the cat, were more likely to produce an arc-like gesture than Japanese speakers who do not have a verb 'swing'. Instead, Japanese speakers often use a straight gesture. Hence the conclusion of the authors that gestures may be shaped by properties of the language of the gesturer.

Gullberg (submitted) also found evidence that the semantics of placement verbs can affect the nature of gesture. In French, 'mettre' ('put') is used as a general placement verb, whereas Dutch has different verbs for putting things somewhere, depending on the type of object that is being put: 'zetten' ('set'), and 'leggen', ('lay'). In the Dutch gestures, the shape of the object was represented by hand shape, whereas in the French gestures, only the path of the placement movement was encoded.

In Figure 2, the Language Architecture is illustrated. The distinguishing property is the interaction between the language structure and the shape of the gestures. In the Kita & Özyürek model, the formulator, doing linguistic computations for speech production, interacts with the processes that determine what is going to be represented in gesture. When the "message generator" is notified that the trajectory information (e.g., an arc-like trajectory in the case of a swing scene) is not "readily verbalizable" (K&O, p. 29), the message generator will drop the trajectory information and adapt the planned gesture accordingly. A Japanese speaker talking about the swing scene will thus produce a straight gesture instead of an arc gesture because there is no motion verb in Japanese that expresses an arc-like trajectory. So by interprocess communication between verbal formulation processes and gesture generation processes it is ensured that the gesture and speech do not reveal conflicting information.

The *Postcard Architecture* implies that information to be communicated is dispatched into gesture and speech channels by a central process. This is presented in De Ruiter (1998, 2000) and is endorsed by Kendon (2004), who writes that "the gestural component of the utterance is under the control of the speaker in the same way as the verbal component, and that it is produced, as spoken phrases are produced, as part of the speaker's *final product*" (p. 156, italics in original). Part of the information to be communicated is expressed in speech, other information in gesture, and although some information will be redundant, and other information will be complementary, gesture and speech are explicitly planned[4] together, to communicate a coherent multimodal message. An utterance is a carefully crafted

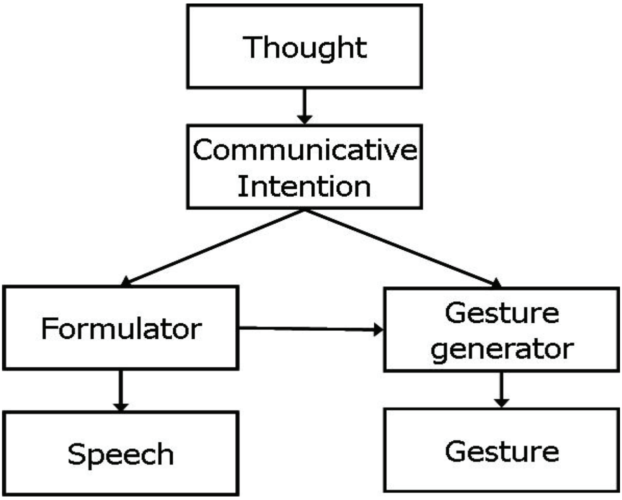

**Figure 2.** The Language Architecture

postcard from the mind, providing the interlocutor with both text (speech) and the accompanying visual illustration (gesture) in the same multimodal message. Figure 3 shows the essential properties of the Postcard Architecture.

In the Postcard Architecture, all the information expressed in gesture and speech is assumed to be communicative, in the sense that it is produced as part of the speaker's communicative intent. That is not to say that every aspect of a gesture is under conscious control, but rather that the function of the speech/gesture system as a whole is to communicate. Melinger and Levelt (2004) provide experimental support for the assumptions underlying the Postcard Architecture. They found that speech that is accompanied by gesture is less explicit than speech that isn't. To explain this distribution of labor between gesture and speech one needs to assume that gesture and speech are planned together at an early stage in utterance production (e.g., the stage that Levelt, 1989, has called the 'conceptualizer'.)

Both the Language Architecture and the Postcard Architecture specify a process that is not explicitly present in the Window Architecture, namely the one labeled 'Gesture Generator'. Transforming a thought (or that part of thought that is to be communicated) into a gesture is a complex process that has been neglected in most work on the processing architecture of gesture. I address this issue in more detail below.

## The Window Architecture and the Postcard Architecture

There are two main arguments for the assumption that gesture is a window into the mind. The first one is the argument that gesture sometimes reveals thoughts that the speaker did not want to share with his interlocutor.

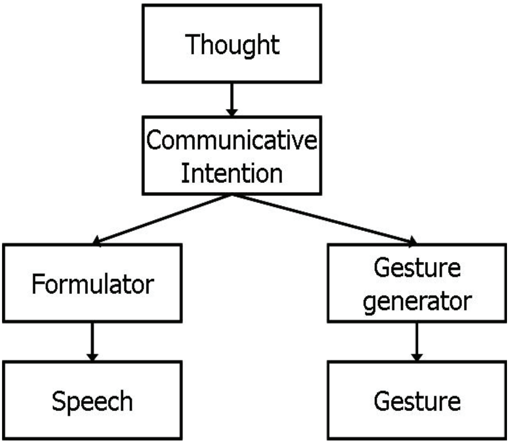

**Figure 3.** The Postcard Architecture

A second argument for the Window Architecture is that gestures are similar to the thoughts they arose from, because a) gesture is not organized in (possibly recursive) syntactic sequences, and b) gesture does not need to conform to lexical and morphological conventions. While these two claims are widely accepted, it does not follow that gestures are more similar to the original thought they express than speech is.

The difference between these two arguments for the Window Architecture is related to what is meant by 'thought'. In the first argument (that gesture production involve less complex transformations) 'thought' refers to processing that is related to the communicative effort (as in 'thinking-for-speaking', see Slobin, 1996, and also McNeill & Duncan, 2000), whereas in the second argument, 'thought' refers to cognitive processes that are not directly related to the communicative intention.

Either way, transforming thought into gesture involves computational processes that are *different* from those related to the production of speech, but this does not imply that they are less complex than for speech. The more complex the transformation from thought to gesture, the larger is the differences between the two. Large differences between a gesture and the thought it expresses (either voluntary or not) is incompatible with the view that gesture gives us a direct window into either concealed or pre-verbal thought.

Let us consider what is computationally (minimally) involved in translating a thought into a gesture.

### *Information selection and perspective assignment*

Transforming a thought into an observable, overt gesture involves several nontrivial computational processes. The first two are very similar to processes assumed to be involved in producing speech, while subsequent processes are very different. These differences in later processing reflect the fact that gesture and speech have different physical properties. To illustrate the processing requirements involved in gesture generation, I will use an instance of the often-quoted 'roll down the hill' speech/gesture fragment.

In this example, a native speaker of English describes the scene from a Sylvester & Tweetybird cartoon of which a still image is presented in Figure 4. In the scene, Sylvester the cat involuntarily swallows a bowling ball and rolls down the street with the ball inside its belly. The example presented here is one of the recorded gestures in the study by Kita & Özyürek (2003).[5] The narrator's speech is

> <u>He rolls down <.></u> the street into a bowling alley

Simultaneously with the underlined part of the transcript above, the narrator makes a spiraling motion with the index finger of the right hand (Figure 5). The gesture

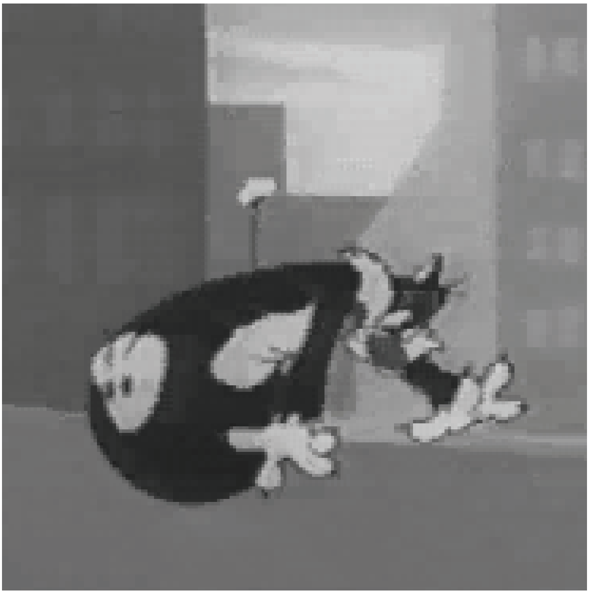

Figure 4. Still from 'Canary Row'

starts slightly before the utterance of 'He rolls down', but is not finished until after 'down' has been produced. The gesture is a spiraling motion of the tip of the index finger, performed in front of the body while the hand moves diagonally downward towards the speaker's right.

Let us now discuss the processing involved in producing gesture, using the gesture described above as an example. First, there is a *selection* process. Even if we assume that thought only consists of propositional representations (e.g., predicates)

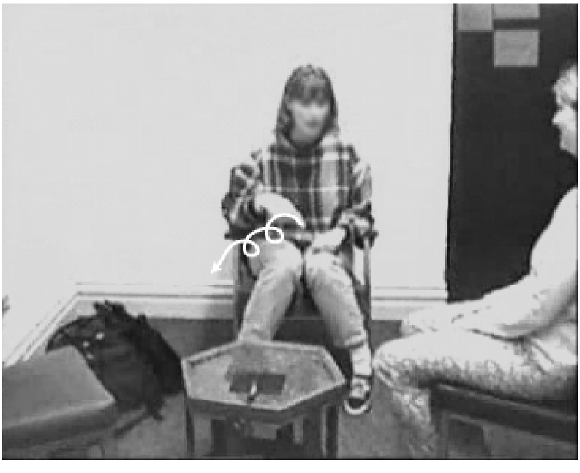

Figure 5. Narrator gesturing 'rolls down the hill'

and imagistic (spatial) representations, and ignore representations in other modalities that might be part of thought, we must assume that there are many more things represented in cognition than are expressed in an utterance. Speakers tend to express information in gesture that is relevant or salient (McNeill, 1992, p. 125). From all the representations that are active in the speaker's mind, only a selected representation (or a specific aspect thereof) is chosen by the speaker to be represented in gesture. Looking at the example, in the original (stimulus) scene, there is not only a distressed cat moving down an alley, but there are also at least three buildings and a street lamp visible. But even if we assume that the narrator does not remember these background details, there is still a cat that frantically moves its paws and legs around, with a nose, eyes, ears, and a distressed facial expression. Nevertheless, the speaker narrating this particular scene produces a gesture that represents the manner and path of the downhill movement of the cat, rather than any other feature of the scene. Note also that in the original stimulus fragment, there is nothing visibly rolling present. In fact, the bowling ball inside the cat cannot be rolling because it is surrounded by cat and therefore not in touch with the street surface. The laws of cartoon physics together with the willing suspension of disbelief lead viewers to infer that something is rolling, and it is this *inference* that is selected for being communicated in the gesture.

This selection process is comparable to the process in speech production that Levelt (1989, p. 123) has called "macroplanning", which is a sub-process of the conceptualizer in Levelt's blueprint for the speaker. In macroplanning, the speaker decides what information is to be expressed in speech. The *Sketch model* by De Ruiter (2000), a Postcard Architecture, assumes that Levelt's conceptualizer is responsible for the selection of information to be expressed for both gesture and speech. Support for this assumption comes from the fact that gesture and speech exhibit semantic and temporal coordination. The information expressed in gesture is semantically related to the speech that is uttered at the same moment (McNeill, 1992; Kendon, 2004). In other words, the speaker does not gesture about one of the buildings in the scene, or imitate the cat's paw movements, while speaking about the motion of the cat. Instead, her gesture expresses the manner and/or path of the movement of the cat down the street, and at the same time in the speech there is the phrase 'rolls down the street'. The temporal/semantic coordination of gesture and speech illustrates that the selection process for gesture is not only important for expressing the most salient and relevant imagistic information, but also for producing the gestural and verbal parts of a communicative act such that they overlap in time, enabling listeners to perceive the two parts of the communicative act as belonging together.

A second computational problem for the speaker is the assignment of viewpoint or *perspective*. In gesturing about the scene in Figure 1, a speaker can select the perspective of the viewer, e.g., by tracing out a trajectory with the index finger from left to right and in front of his body, or they can take the perspective of the cat, by tracing the trajectory forward and downward from the center of his body (for more detailed discussion of perspective taking in gesture, see McNeill, 1992, and Levinson, 2003). There is a parallel process in the production of speech (Ehrich, 1982; Levelt, 1984; Linde & Labov, 1975) which in Levelt's (1989) architecture is called 'microplanning'. As McNeill (1992) has shown, the perspectives taken by the speaker in gesture and speech are highly correlated. Again, this argues for a single computational process that plans gesture and speech together. In the example, the speaker has chosen the cat-external perspective of the viewer. We know this because in the cartoon, the cat moves downwards and from left to right, and the gesturing hand moves in the same direction and with roughly the same downward slope (from the speaker's perspective).

### *Gesture generation*

The selection of information to be expressed and the assignment of a perspective for the expression of that information are structurally similar processes for gesture and speech. However, after a speaker has selected the relevant information to be expressed, and assigned a perspective, the subsequent processing of gesture and speech will be very different. In speech, syntactic structures and lexical forms need to be selected, morphosyntactic agreement has to be computed, and the resulting representation must be transformed in a series of articulatory gestures for the generation of overt speech. This sounds like an impressive amount of computational work, and indeed it is. But generating a gesture also involves a series of nontrivial computations.

Gesture generation transforms the selected information (with the assigned perspective) into an overt, observable (physical) gesture. The problem of generating an overt gesture from an abstract (presumably spatio-temporal) representation is one of the great puzzles of human gesture, and has received little attention in the literature. In De Ruiter (2000) I sketch a possible processing mechanism for gesture generation based on a constraint satisfaction mechanism, involving successive reduction of motoric degrees of freedom. This approach has been also used in Kopp (2003), who implemented a gesture/speech production module for use in virtual agents. Alternatively, Kita & Özyürek (2003) propose that gestures are generated "on the basis of action schemata which are selected on the basis of features of imagined or real space" (p. 28). However, it is an open question whether for all gestures a pre-existing action schema can be identified, or whether there are gestures (esp. the ones that are not pantomimes) that rely on newly generated motion patterns created ad-hoc for gestural communication.

The most intriguing computational problem concerns 'audience design' (Clark & Carlson, 1982), or 'recipient design' (Sacks & Schegloff, 1979). For both speech and gesture, the speaker's problem is how a communicative action is to be performed such that the addressee(s) can be expected to be able to interpret its meaning. The fact that gesture is not governed by strictly conventional form-meaning mappings (as is the case for words in spoken languages) makes this problem harder for gestures than it is for speech. Another intriguing aspect of recipient design in gesture becomes apparent when speakers address more than one listener. Özyürek (2002) shows that speakers adapt their gestures to the location(s) of their interlocutor(s). Speakers also take into account which communicative modalities are available to their listeners. Bavelas et al. (2002) show that speakers who were aware that their descriptions would be seen on video made more gestures than when they thought their description would only be heard.

Recipient design in gesture does not come for free. Because listeners are so good at comprehending gesture (when they also have access to the accompanying speech), we might feel that this is a simple process. But to date there still is no intelligent (AI) system that can spontaneously[6] generate iconic gestures that make sense to human observers. In contrast, we do have many computer systems that spontaneously generate synthetic speech that we can understand. Although speech production involves a lot of processing, we do have a reasonable level of understanding of the *nature* of the processing involved. For gesture generation, we don't have the same level of understanding, although there are suggestions in the work of Calbris (1990) of how gestures could be built up from a pre-existing repertoire of gestural elements. Also, Kopp (2003) shows significant progress in modeling the generation of motor programs from abstract gestural specifications.

For the 'roll down the hill' example described above, I assume that the gesture is generated as follows: As human hands cannot literally perform a rolling motion, the subject's solution is to take an arbitrary point on the surface of the imagined bowling ball, and use her fingertip to trace out in the space in front of her the movement that this imaginary point makes. The rotating index finger, combined with the rightward and downward motion of the hand, are communicating the 'rolling down' concept expressed in the speech. The gesture also conveys information not expressed in the speech, namely the speed, angle and direction of the rolling down motion.

A further computational problem to be solved is *hand allocation*. Are both hands free (i.e., not engaged in another activity)? Is the gesture going to be performed with the left or right hand, or with both? What are the 'roles' assigned to each hand? Sometimes, one gesture takes on the function of a place marker for a previous gesture, while the other hand performs a new gesture (Enfield, 2004).

In the example, the subject chose to use the right hand, which may be because it is her dominant hand.

Finally, the gesture generation process needs to take into account *environmental constraints* (De Ruiter, 2000). The gesturer needs to prevent, for instance, hitting someone in the face or knocking over a glass while performing a gesture.

The above description of the computational processes involved in gesture production is not intended to be exhaustive. I hope to have shown that the end product of gesture production (i.e., an observable gesture) and its origin (a thought) are very different entities, and that a gesture is not a literal image of the thought that it expresses. The properties of gestures mentioned above strongly suggest that gestures and speech are designed from the earliest moment to form a coherent and interpretable multimodal signal.

### *Consciousness, intentionality, and communicativeness*

A second argument that gesture is a window into the mind is that it often reveals information about the speaker's thought that was not consciously intended to be communicated (see the quotes of Beattie and McNeill above). I contend that *most* of the communicative signals that we produce in interaction are not consciously planned, and this holds for speech as well as for gesture. The thought that is to be expressed is probably in our consciousness, but most aspects of the way we express this thought are governed by automatic behavior not under our conscious control. In producing speech, for instance, speakers choose a certain prosody (rhythm, intonation, amplitude envelope), voice quality (angry, neutral, sweet, pleading), syntax (choosing from a multitude of possible syntactic structures, each with consequences for the order in which the information is expressed), and semantics (which word is used to express a certain concept). Speakers cannot be consciously making decisions about all these aspects of speech; it is the largely automated nature of the speaking process that enables us to speak so fast and fluently (Levelt, 1989). Similarly gesture, being a largely spontaneous activity, will not be under total conscious control. This does not mean that gesture is not intended to communicate exactly that which the speaker wishes to communicate.

### *Gesture-speech 'mismatches' in children*

Goldin-Meadow & colleagues (Alibali & Goldin-Meadow, 1993, Church & Goldin-Meadow, 1986, Goldin-Meadow et al., 1993) have discovered that children sometimes produce gestures that, in the terminology by Goldin-Meadow et al., 'mismatch' with the accompanying speech. A mismatch means that the gesture "contained different information […] from that contained in the speech" (Church &

Goldin-Meadow, 1986, p. 53). So a mismatch occurs not only when the information in gesture and speech are contradictory, but also when they are just different (p. 53). Because these gesture/speech mismatches can predict whether a child is about to reach a new level of understanding in cognitive tasks, the gesture (together with the speech) could be said to provide a window into the (developing) mind.

A problem with this definition of a mismatch is that imagistic gestures *always* reveal information that is different from the information expressed in the affiliated speech. In the 'roll down the hill' example discussed above, an example that is often quoted as a prime example of semantic synchrony between gesture and speech, the speed, angle and direction of the gesture were present in the gesture, but not in the speech. In the definition by Church and Goldin-Meadow, however, this would nevertheless be a *mismatching* speech/gesture combination.

It is an intriguing question how the contradictory mismatches arise. One possibility is that the ability to produce semantically or pragmatically congruent gesture/speech combination is a skill that is not 'built-in' the speech/gesture system, but has to be acquired (Goldin-Meadow, personal communication). This predicts that the frequency of contradictory mismatches will decline with age.

### *Summary*

Gestures are not direct, 'raw' representations of unformulated thought. They are rather carefully crafted visual messages designed to be understood in combination with the accompanying speech. For gestures to be recognizable as gestures, and to be understandable for interlocutors, they have to be designed to accomplish that. Bavelas et al. (2002), Melinger & Levelt (2004), and Özyürek (2002), among others, provide evidence that speakers design their gestures to accommodate listeners. We cannot assume that gesture is a window into the mind, just as we cannot assume that language is.

## The Language Architecture and the Postcard Architecture

The Language Architecture is different from the Window Architecture in that it does not assume that gesture is a window into the mind. Rather, it assigns a prominent role to the language processing involved in speaking and gesturing. Gestures are partly shaped by the (analog) features in the original representation, and partly by the representations active in speech. Kita & Özyürek (2003) present evidence for structural aspects of languages shaping the gestures of native speakers of those languages. They propose a gesture/speech model to accommodate their findings. Their model assumes communication between the formulator and other processes involved in modality selection and message generation (see Kita & Özyürek, Figure 7 and accompanying text for details). If the formulator finds out that it cannot encode the message that it receives properly it will send feedback to the message generator which then creates a new message (or series of messages) to accommodate. A crucial example is the narration of the scene from the Sylvester & Tweetybird cartoon where Sylvester the cat 'swings across' a street using a rope of some sort. As Japanese and Turkish do not have a lexical concept that means 'swing', many (but not all) Japanese and Turkish speakers produce a straight gesture when describing the path of the movement by the cat, whereas English speakers who *do* have the verb 'swing' in their language, generally produce arc-like gestures.

The Postcard Architecture can also accommodate language-specific effects on gesturing, but in a way that is fundamentally different from the Language Architecture. In the Postcard Architecture, the fact that in Japanese or Turkish there is no concept with the same meaning as English 'to swing', is known by the process responsible for encoding the communicative intention. Hence, it will either generate a straight gesture that is compatible with other motion verbs such as 'move', or an arc-like gesture which is still consistent with the verbal message, but gives more analog information about the path. This depends on whether the speaker wants to communicate the path information at that level of detail. The Language Architecture assumes that properties of the language result in interprocess communication to resolve discrepancies between gesture and speech. In the Postcard Architecture, these discrepancies do not arise because the process that plans the (multimodal) utterance has access to the concepts that are available in language (cf. Roelofs, 1992).

It is possible to derive different and testable predictions from the Sketch model by De Ruiter (2000) — a postcard architecture — and the Kita & Özyürek model — a language architecture — for these specific cross-cultural comparisons. In the Kita & Özyürek model, for the case of Japanese and Turkish speakers wanting to speak and gesture about the swing-scene, the message generation processes will receive a message from the formulator that "the trajectory shape is not readily verbalizable" (Kita & Özyürek, p. 29) The Message Generator, upon being informed of this, will generate a new gesture that leaves out the trajectory shape information. In contrast, the Message Generation process in the Sketch model will already know beforehand that the trajectory information cannot be expressed in speech, and will have either put this information in gesture or not, depending on whether this level of detail is required for the communicative intention. A differential prediction between these two accounts is one of timing. In the Language Architecture, there will be, sequentially, a) a message to the formulator, b) feedback from the formulator, c) adaptation of the message, and d) a new message to the formulator and gesture generator. The Sketch model does not need extra computational cycles to encode the swing scene in gesture and speech. If the Kita & Özyürek model is correct (and the Sketch model is not) then we would expect to see a timing difference between Japanese and Turkish speakers narrating about a 'swing' scene on the one hand, and a 'move' (any motion trajectory that is expressible in the language will do here) scene on the other. The 'swing' scene descriptions should be produced with a greater delay than the 'move' scene descriptions. If the Sketch model is correct (and the Kita & Özyürek model is not) then we do not expect this timing difference. Further experimentation and/or re-inspection of the data collected by Kita & Özyürek could shed more light on this issue.

## The Window Architecture and the Language Architecture

If the structure of language can influence the shape of gestures, as Kita & Özyürek (2003) show, this is in stark contradiction with the Window Architecture. The whole point of the Window Architecture is that linguistic processing is *bypassed*, which is why it provides us with a window into the mind. A possible defense of the Window Architecture is to argue that some gestures provides us with a window into the mind, whereas others, such as the ones observed by Kita & Özyürek, do not. This weaker claim would be very hard to falsify, as it is seems almost impossible to establish, in general, whether a certain gesture has been influenced by the structure of the speaker's language or not.

## Summary and conclusion

Imagistic gestures are not literal images of a speaker's thoughts. Speakers select thoughts to be communicated, and in order to achieve communicative success, gestures need to be designed to accomplish this. In other words, gestures need to be recipient-designed together with the accompanying speech. Gesture informs us about the thoughts of the speaker, but so does speech. The analogous statement that 'speech is a window into the mind', would be either trivial, in the sense that we obviously gain information about the speaker's mind from their speech, or very wrong, in the light of the complex processing necessary to transform a communicative intention into articulatory behavior. The transformation of a thought into an overt gesture is different from, but not necessarily less complex than, the processes that transform communicative intentions into speech, and that these transformations prevent gesture from being a window into the mind. The fact that listeners can interpret gestures with relative ease (if they have access to the speech as well) is precisely why they cannot be windows into the mind. Both the Language and Postcard architectures incorporate this assumption, but they differ in the way the semantic synchronization between language and gesture is achieved. Further research into the relative timing of gesture and speech could help us gain more insight into how gestures and speech are orchestrated together to form coherent multimodal communicative actions.

## Notes

[1] For a definition of a functional communicative modality, see De Ruiter et al. (2003).

[2] As I mention below in the discussion of the study by Kita & Özyurek, not all gesture researchers believe that imagistic gestures have a completely idiosyncratic form-meaning relationship. Gestures could also be composed from a standard repertoire of representational techniques (see also Calbris, 1990 and Kendon, 2004, ch. 9).

[3] To avoid confusion: sign language *does* have morpho-syntactic structure, but we are limiting the discussion here to the spontaneous gestures that accompany spoken language.

[4] The word 'planned' does not imply that this process is neccessarily a conscious process. This will be discussed below.

[5] The author wishes to thank Asli Özyürek and Sotaro Kita for generously supplying the video fragment of this example.

[6] By spontaneously I mean that the generated gesture is created anew from the communicative intention and the information present in thought at that moment, as opposed to retrieved from some kind of gestural 'lexicon.'

## References

Alibali, Martha & Susan Goldin-Meadow (1993). Gesture-speech mismatch and mechanisms of learning: What the hands reveal about a child's state of mind. *Cognitive Psychology*, 25, 468–523.

Bavelas, Janet, Christine Kenwood, Trudy Johnson , & Bruce Phillips (2002). An experimental study of when and how speakers use gestures to communicate. *Gesture*, 2 (1), 1–17.

Beattie, Geoffrey (2003). *Visible thought: The new psychology of body language*. Hove, UK: Routledge.

Calbris, Geneviève (1990). *The semiotics of French gestures*. Bloomington: Indiana University Press.

Church, R. Breckinridge & Susan Goldin-Meadow (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition*, 23, 43–71.

Clark, Herbert H. & Thomas B. Carlson (1982). Hearers and speech acts. *Language* 58, 332–373.

Ehrich, Veronika (1982). The structure of living space descriptions. In R. J. Jarvella & W. Klein (Eds.), *Speech, place, and action. Studies in deixis and related topics* (pp. 219–249). Chichester: John Wiley.

Enfield, Nick J. (2004). On linear segementation and combinatorics in co-speech gesture: a symmetry-dominance construction in Lao fish trap descriptions. *Semiotica*, 149 (1/4), 57–123.

Enfield, Nick J. (2005). The body as a cognitive artifact in kinship representations. *Current Anthropology*, 46 (1), 51–81.

Goldin-Meadow, Susan, Martha Alibali , & R. Breckinridge Church (1993). Transitions in concept acquisition: Using the hand to read the mind. *Psychological Review*, 100, 279–297.

Gullberg, Marianne (submitted). Language-specific encoding of placement events in gestures. In E. Pederson, R. Tomlin, & J. Bohnemeyer (Eds.), *Event representations in language and cognition*.

Kendon, Adam (1972). Some relationships between body motion and speech. In A. W. Sigman & B. Pope (Eds.), *Studies in dyadic communication* (pp. 177–210). New York: Pergamon Press.

Kendon, Adam (1980). Gesticulation and speech: Two aspects of the process of utterance. In Mary R. Key (Ed.), *The relatinship of verbal and nonverbal communication* (pp. 207–227). The Hague: Mouton.

Kendon, Adam (1988). *Sign languages of aboriginal Australia: Cultural, semiotic, and communicative perspectives*. Cambridge: Cambridge University Press.

Kendon, Adam (1990). Gesticulation, quotable gestures and signs. In M. Moerman & M. Nomura (Eds.), *Culture embodied* (Vol. 27, pp. 53–77). Osaka: National Museum of Ethnology.

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

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

Kendon, Adam & Laura Versante (2003). Pointing by hand in 'Neapolitan'. In S. Kita (Ed.), *Pointing: Where language, culture and cognition meet* (pp. 109–137). Hillsdale, NJ: Lawrence Erlbaum.

Kita, Sotaro & Asli Özyürek (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48, 16–32.

Kopp, Stefan (2003). *Synthese und Koordination von Sprache und Gestik für virtuelle multimodale Agenten*. Berlin: Akademische Verlagsgesellschaft Aka GmbH.

Levelt, Willem J. M. (1984). Some perceptual limitations on talking about space. In A. J. Van Doorn, W. A. Van der Grind, & J. J. Koenderink (Eds.), *Limits in perception* (pp. 323–358). Utrecht: VNU Science Press.

Levelt, Willem J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: The MIT Press.

Levinson, Stephen C. (2003). *Space in language and cognition: Explorations in cognitive diversity*. Cambridge: Cambridge University Press.

Linde, Charlotte & William Labov (1975). Spatial networks as a site for the study of language and thought. *Language*, 51, 924–939.

McNeill, David (1992). *Hand and mind*. Chicago, London: The Chicago University Press.

McNeill, David (Ed.) (2000). *Language and gesture: Window into thought and action*. Cambridge: Cambridge University Press.

McNeill, David & Susan Duncan (2000). Growth points in thinking-for-speaking. In D. McNeill (Ed.), *Language and gesture* (pp. 141–161). Cambridge: Cambridge University Press.

Melinger, Alissa & Willem J. M. Levelt (2004). Gesture and the communicative intention of the speaker. *Gesture*, 4 (2), 119–141.

Özyürek, Asli (2002). Do speakers design their cospeech gestures for their addressees? The effects of addressee location on representational gestures. *Journal of Memory and Language,* 46, 688–704.

Roelofs, Ardi (1992). A spreading activation theory of lemma retrieval in speaking. *Cognition*, 42, 107–142.

De Ruiter, Jan Peter (1998). *Gesture and speech production.* Unpublished Dissertation, Nijmegen.

De Ruiter, Jan Peter (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture* (pp. 284–311). Cambridge, UK: Cambridge University Press.

De Ruiter, Jan Peter (2003). The function of hand gesture in spoken conversation. In M. Bickenbach, A. Klappert , & H. Pompe (Eds.), *Manus loquens*. Cologne: Dumont.

De Ruiter, Jan Peter, Stéphane Rossignol, Louis Vuurpijl, Douglas C. Cunningham , & Willem J. M. Levelt (2003). SLOT: A research platform for investigating multimodal communication. *Behavior research methods, instruments, & computers*, 35 (3), 408–419.

Sacks, Harvey & Emmanuel A. Schegloff (1979). Two preferences in the organization of reference to persons in conversation and their interaction. In G. Psathas (Ed.), *Everyday language: Studies in ethnomethodology* (pp. 15–21). New York: Irvington.

Slobin, Daniel (1996). From "thought and language" to "thinking for speaking". In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70–96). Cambridge: Cambridge University Press.

Wilkins, David P. (2003). Why pointing with the index finger is not a universal (in sociocultural and semiotic terms). In S. Kita (Ed.), *Pointing: Where language, culture and cognition meet* (pp. 171–215). Hillsdale, NJ: Lawrence Erlbaum.

## *Author's address*

Jan Peter de Ruiter
Max Planck Institute for Psycholinguistics
P.O. Box 310
NL-6500 AH NIJMEGEN
The Netherlands

E-mail: janpeter.deruiter@mpi.nl

## *About the author*

**Jan Peter de Ruiter** is a scientific staff member in the Language and Cognition Group of the Max Planck Institute for Psycholinguistics, and is also involved in several multidisciplinary research projects funded by the European Union. His main interests are in multimodal human-human and human-machine communication, gesture, mathematical modeling of human behavior, and turn-taking. Publications include *The production of gesture and speech* (Cambridge University Press, 2000); Can gesticulation help aphasic people speak, or rather, communicate? *Advances in Speech-Language Pathology*, 2006; and: Predicting the end of a speaker's turn; a cognitive cornerstone of conversation. *Language*, 2006.