ABSTRACT

This paper discusses some of the major psycholinguistic theories of coordination between speech and gesture proposed in the recent years. Among the theories of speech-gesture coordination to be discussed here are two modular theories (Lexical Retrieval model and Sketch model), two interaction theories (Growth Point theory and Information Packaging Hypothesis), and yet another hypothesis that gesture derives from instrumental action. In discussing these different theories, we clarify the nature of the issues and try to determine which theoretical approach is most promising.

Statement of the Problem

The term speech is employed here in the sense of an actual use of language, i.e., Saussurian parole as opposed to langue. The term gesture refers mainly, but not exclusively, to a manual sign which in and of itself is not speech, but typically accompanies it. Some gestures are called emblems. They presuppose symbolic and hence conventional form-meaning relationship, of which an “okay” sign is an example. A pointing gesture presupposes some conventional form-meaning relationship to the extent that it has a semantic definition. It also depends on the context of use to the extent that the referent(s) cannot by definition be determined without observing when and where a token is produced. Other gestures do not presuppose any conventional regularity: they are merely iconic indices at best. The meaningfulness of such a gesture comes from the iconicity (i.e., likeness) and/or the indexicality (i.e., spatiotemporal contiguity) between the form and its semiotic object strictly in Peirce’s sense of the terms. Such a gesture is meaningful through the contrasts created, in the context of its particular moment of use, by its iconic and indexical characteristics.

Gesture in and of itself is not speech, regardless of whether it presupposes a conventional form-meaning relationship. This is because it does not presuppose the syntagmatic and paradigmatic axes upon which speech is conventionally regulated. In this sense, gesture should be distinguished from a sign language, such as American Sign Language, even though they share the same visuo-
A Review of Theories of Speech-Gesture Coordination

spatial modality. Non-conventional (iconic and/or indexical) gestures and pointing gestures typically accompany speech, while emblems (symbolic/conventional gestures) do so less typically.

Speech and gesture are said to express similar ideas or different but often complementary aspects of one and the same state of affairs and/or ideas. (See, however, Goldin-Meadow, Alibali and Church’s study[4] and the discussion below of a speech-gesture mismatch phenomenon.) What motivates us to take co-occurring speech and gesture as expressing largely synonymous ideas is the fact that they are in a mutually indexical relationship.

How they index each other is rather an empirical question. Some gestures seem to synchronize with the meaningfully related speech segments, while others seem to either precede or follow such speech segments (e.g., a pointing gesture preceding/following a speech segment like, ‘Look at that!’). How to explain such data is a major theme of research. The goal of the present paper, then, is to review some of the major psycholinguistic theories of speech-gesture coordination that have been proposed in the recent years. In so doing, we attempt to clarify the nature of the issues and determine which theoretical approach is most promising.

2 Non-Interactive Information Processing model

This section discusses two theoretical models of speech-gesture coordination. One is the Lexical Retrieval model,[6][7] and the other is the Sketch model. Both are based on Willem Levelt’s Information Processing model of speech production (or IP model, for short),[5] which I will briefly introduce.

According to the IP model, there are several modules in the speech production system, called the Conceptualizer, Formulator, and Articulator, in the order of the flow of information. Lemma, the storage of lexical items, is accessible from the two sub-modules of the Formulator; i.e., the grammatical encoder and the phonological encoder. It is claimed that these modules do not interact with each other. There are some feedback loops, and they are the only way the information can be sent back to the starting point of the process. The whole process starts off with preparing a preverbal message which, as the name suggests, is not yet encoded linguistically. The message is thus assumed to be amodal, merely propositional, and universal. It is a pure idea in the sense that it is not affected by the structural characteristics of any particular language. The output of the Conceptualizer, i.e., a preverbal message, can only enter the next module, where it is transformed into a verbal message. Thus, in the Formulator, the necessary lexical items are retrieved from the Lemma to denote each component of the proposition. The lexical items are then grammatically encoded to reflect the internal relationship among the components of the preverbal message. The output of the phonological encoder is a phonetic plan of the sentence, which then enters the Articulator where the phonetic plan is transformed into a motor plan to be finally executed by the articulatory organs.

2.1 Lexical Retrieval Model (LR model):

Based on Levelt’s IP model of speech production, the Lexical Retrieval Model (or, LR model for short) makes two basic arguments regarding the function of gesture:

(1) Gesture is not communicative, and is not even intended by the gesturer.[6][7] This argument is based on the observations that (i) people cannot determine the meaning of a gesture when the videotaped image of it is played back with the sound channel off, i.e., without the speech with which the gesture originally co-occurred, and (ii) people gesture even when there is no visual contact with the interlocutor; e.g., when speaking on the telephone.

(2) Gesture helps one speak fluently and, in particular, helps retrieve lexical items the speaker has trouble retrieving.[6][7] This argument requires a specific assumption that lexical items are stored in the long-term memory such that each of them is associated with a visual image of the referent(s). Gesticulation activates the visual image, which then activates the target lexical item. The only available evidence is indirect at best, and it basically could show either of the following two things: one is that people gesture more often when they have a problem with retrieving a lexical item than otherwise. The other evidence is that patients with a certain damage and/or anomaly of the brain who are diagnosed as having memory problems gesture more often than those who are normal in that regard.

A couple of comments are in order. First, the propo-
ponents of the LR model seem to assume that for a sign to be communicative, its referent(s) should only be determinable by making reference to the conventional code, according to which the message is encoded. In other words, only symbols count as communicative. With this criterion, we do not even have to conduct an experiment, since the gestures under discussion here include non-symbolic ones which are not “communicative” by definition. There is a problem with the assumption that only pure symbols count as communicative, however. If we accept this assumption, we also have to accept that demonstrative expressions (e.g., that) are, strictly speaking, not communicative either. This is because we cannot determine the referent(s) of a demonstrative expression without looking at the situation in which such an expression is used. Gesture is communicative when taken together with the concurrent speech just as demonstrative expressions are communicative when taken together with the context of their use.

![Figure 1: ASL alphabet G.](image)

Secondly, the claim that gesture helps lexical retrieval, besides having no direct evidence, has an internal flaw as a realistic psychological mechanism. That is, visual images that could possibly be activated by a gesture would be infinitely many. Consider, for instance, how many things one can talk about with a G-hand (see Figure 1 above), the hand shape that is often used for pointing. For some speakers, it is a cat. For others, the tip of the index finger is a molecule. For still others, it is a pointer! And so on and so forth. The visual images that can possibly be invoked by a gesture should be too ambiguous to be in the service of helping a lexical retrieval. Related to this is the fact that people seem to gesture when they have problems retrieving lexical items for an abstract concept, in pretty much the same way as when the target lexical items are for a concrete object. Besides, gestures produced when the speaker has a lexical retrieval problem apparently are, in many cases, beats, which would not invoke any “visual image” at all. The LR model does not appear to be convincing enough despite the intuitively appealing arguments.

### 2.2 Sketch Model

In contrast with the LR model, the Sketch model, another IP model of speech and gesture, argues that gesture is communicative. It also argues that, while rejecting the possibility that gesture might help lexical retrieval, “gesturing facilitates access to the representations in memory that the gesture is generated from.” According to the Sketch model, the information necessary for producing a gesture and the concurrent speech is prepared in the Conceptualizer. Levelt’s Conceptualizer is revised such that it can accommodate a mechanism which splits the conceptual information into two parts: one is a preverbal message and the other is information called “Sketch.” The preverbal message is exactly the same as the one in Levelt’s model. The Sketch is the information that is eventually to be gesturally represented. The preverbal message enters the Formulator to be grammatically encoded, and the Sketch enters the Gesture Planner to be formulated into a motor command. As a typical IP model, no interaction between the modules are allowed except through the feedback loop(s). The evidence for modularity comes from an experiment on pointing. Even when gesticulation of a pointing movement is distracted by means of physically immobilizing the speaker’s hands and/or arms (e.g., by the experimenter grabbing them) at the moment of executing a motor command for a preplanned gesture, the supposedly concurrent speech is not distracted, and the articulation is often completed without the gesture. Thus, it is argued, once the planning of a sentence and a gesture is completed in the Conceptualizer, the grammatical encoding, the translation into a motor plan, and the articulation are automatic and independent from the formulation and execution of a gesture. Hence, the Sketch model argues for the ballistic model of a pointing gesture.

In a series of experiments conducted by De Ruiter the rate of gesture occurrence per word is shown to be higher when the speaker describes a spatial layout from memory than when doing the same thing with a picture of the layout visible to her/him. This is taken as supporting De Ruiter’s version of the Retrieval Hypothesis that “gestur-
A Review of Theories of Speech-Gesture Coordination

...ing facilitates access to the representations in memory that the gesture is generated from,” and not the lexical retrieval as such. This makes it easier, it is argued, for the speaker to think about the visuo-spatial representation, and to produce a preverbal message about it, because the representation is active and immediately available in the working memory. Furthermore, in the same series of experiments, the rate of gesture occurrence per word was found to be no less frequent when the layout is complex and difficult to describe than when it is simple and easy to do so. It is further argued that this disproves the lexical retrieval hypothesis of the LR model, which proposes that gesturing facilitates the process of speaking by way of helping lexical retrieval. This interpretation may not be so obvious, and requires the assumption that the more complex spatial layout requires more infrequent, and thus more difficult to retrieve lexical items than the description of the less complex one. The finding that there was no difference in the amount of gesture between the two conditions is thus taken as suggesting that the LR model does not hold. If this in fact is the case, the data regarding the assumption should also be presented. What seems to be difficult for the speaker in this experimental setting is to find out what expression best fits the spatial layout, for which there may not necessarily be a lexical retrieval problem per se. The Sketch model also argues that the finding disproves Kita’s version of the Interaction Model (to be discussed below) that spatio-kinesic thinking and analytical thinking interact with each other and together constitute thought. De Ruiter’s assumption is that the more difficult spatial layout requires more grammatical encoding. The finding that there was no difference in the rate of gesture occurrence per word between the two conditions is thus claimed to suggest that gesture is not used to help this assumedly more complex grammatical encoding. To support the assumption of the argument, the average number of words used for the whole description of the complex layout is reported to be higher than that for the description of the less complex layout. This does not suggest that the grammatical encoding per sentence was also more complex, however. Accordingly, the finding cannot tell us anything about the relationship between the rate of gesture occurrence and the grammatical complexity of the sentences uttered.

We now examine a few more points about the Sketch model. First, it should be obvious that the Sketch model is self-contradictory. This is because the Sketch model argues that the Conceptualizer is equipped with a mechanism that divides raw information into two parts, one to be represented linguistically and the other to be represented gesturally. The Sketch model, however, assumes that the Conceptualizer is not affected by anything linguistic, i.e., the grammatical and lexical categories. But then, how can the Conceptualizer achieve the split of the information for which the linguistic categorial knowledge is necessary, yet absent? An interaction between gesture and speech which is asserted in the Sketch model to be “only possible in an early stage (in the conceptualizer)” necessarily introduces the grammatical categories into the Conceptualizer, thus violating the central doctrine of the modular theory on which the Sketch model bases itself.

The second comment to be made is on the Sketch model’s argument against interaction models: according to the Sketch model, the grammatical encoding is done only in the Formulator. Thus, note that the Conceptualizer, not to mention the Gesture Planner, does not and cannot “know” what lexical items will be picked up and how they are arranged in the surface sentence. This is because the lexical items are selected and arranged in the Formulator, to which the Conceptualizer and the Gesture Planner are not privileged to access. This means that neither the Conceptualizer nor the Gesture Planner can make any plan regarding when the gesture should be executed with respect to the concurrent speech. Gestures seem to occur, however, not randomly but very systematically in terms of the timing with the concurrent speech and of the discursive context, as shown by many studies.

Contrary to the Sketch model’s claim that the gesture should not be affected by the grammar of the language the speaker speaks because the Gesture Planner and the Formulator are independent from each other, there is rather direct evidence that suggests that this is not the case. We will now turn to it.

3 Interaction Models

3.1 A counter example to the non-interaction view

Interaction models in general argue that the organization of speech and gesture are interactive at least from the be-
ginnng up to the Articulator.\textsuperscript{[11]} The celebrated example comes from a Japanese narration of ‘the swing scene’ in a cartoon story.\textsuperscript{[10][11]} This example shows that the linguistic limitation (the lack of a straightforward lexical item in Japanese) is in fact reflected in the way the information splits into two parts, one which is to be represented linguistically and the other gesturally. The speakers try to describe a motion event from a cartoon story in which one of the cartoon characters “swings across the street on a rope that we must imagine attached somewhere in the air above the street” just as Tarzan does. (See Figure 2 for the scene described by the speakers.\textsuperscript{[11]})

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{swing_scene.png}
\caption{The scene described in Kita’s experiment, reproduced here by the present author.\textsuperscript{[11]}}
\end{figure}

The lack of the intransitive verb in Japanese that encodes the information of the arc trajectory of the character’s agentive action led the Japanese speakers to either linguistically or gesturally express the arc trajectory.\textsuperscript{1} Interestingly, ‘68\% (14) of the arc gestures ... were not produced with linguistic encoding of an arc. Thus, they cannot originate from a lexical item in the concurrent speech.’\textsuperscript{[10]} It was also found that ‘38\% of all the Japanese Swing Scene gestures’\textsuperscript{[10]} were “a straight gesture, which does not encode an arc.”\textsuperscript{[10]} Crucially, “it is not the case that Japanese speakers sometimes failed to recall the trajectory. Every Japanese speaker encoded the arc trajectory either linguistically or gesturally somewhere in his/her description ... The arc trajectory was somewhere in the Japanese speakers’ mind, yet they sometimes produced a straight gesture stroke, which does not encode an arc. How could such straight gestures be generated\textsuperscript{[10]} if the content of gestural representation derives solely from the visual memory and/or representation of the swing scene? As Kita pointed out, “The straight gestures of the Japanese speakers show that a gesture is not simply a retrieval of visual stimulus, nor photographic reconstruction of an event. The shape of gesture can change depending on the language even if the change would decrease the iconicity. The decrease in iconicity correlates with the inaccessibility of linguistic encoding of a particular feature of the described event.”\textsuperscript{[10]} The arc gesture without the linguistic encoding of the arc trajectory and the straight gesture despite the active representation of the arc trajectory both point to the following: “Construction of a message crucially involves an interaction between language and imagistic thought.”\textsuperscript{[10]} That is, the origin of gesture is both linguistic and imagistic. By the same token, the origin of speech is both linguistic and imagistic.\textsuperscript{2}

\subsection{Growth Point (GP) theory}

Growth Point theory (or “GP theory,” for short) considers gestures, and speech as well, as “material carriers of thinking”\textsuperscript{[14]} (emphasis in original). To gloss this, speech and gesture can be thought of as the joint embodiment of the production process itself, and not as the packaged communicative output of a separate internal production process.\textsuperscript{[14]} By the “embodiment” of the production process, it is meant that the core idea is brought into concrete existence and becomes part of the speaker’s own existence at that moment.\textsuperscript{[14]} Accordingly, gestures together with the concurrent speech are themselves ‘thinking’ in one of its many forms — not merely expressions of what has already been thought of, but the thinking process itself.\textsuperscript{[14]} By implication, it is not the underlying mental construction alone, but the actual rendering of speech and gesture in the communicative event that completes the meaningfulness of these signs. Translating this into Peircean semiotic terms, it is the indexicality of signs, i.e., the actual event or happening of signs, that achieves the meaningfulness of the signs without which thinking cannot even proceed.\textsuperscript{[3]} This is, as claimed by GP theory, \emph{why} people gesture and speak. It makes much more sense than saying that “gesturing facilitates access to the representations in memory that the gesture is generated from.”\textsuperscript{[3]} This is because, if gestures and speech are material carriers of thinking, thinking would not proceed without them being actually produced. The reason why the Sketch model makes less sense is that it does not explain why gesture has to be \emph{actually} produced, as opposed to being merely imagined, in order for the visuospatial
representation to be activated in memory. One might ask why people do not always gesture if gestures are indeed material carriers of thinking. To be sure, gestures are not the only material carrier, and, in fact, there are some other ways in which thinking is materially carried out, of which an example is writing out one’s ideas[14] or the use of prosody in speech. That should explain why people do not always gesture. The material carrier analysis of speech and gesture in itself, however, does not seem to answer the question why people gesture while speaking. To answer this question, we have to take a look at the central part of Growth Point theory, to which we now turn.

The growth point is a unit of analysis consisting of imagery and linguistic category. It is assumed to be a minimal psychological unit in L.S. Vygotsky’s sense[15] and, as such, retains the essential properties of a whole of an image and a linguistically-codified meaning category.[14] In other words, in a GP, imagery and linguistic category are mediated and, thus, affected by each other from the very beginning of the entire production of speech and gesture through the end of it. Accordingly, a GP has the characteristics of both imagery and language. Its essential feature is a dialectic of these two different forms of thinking.

Speech, as considered in GP theory, indexically (pragmatically) presupposes or entails that a certain gesture is a sign that ‘refers’ to a particular entity (e.g., a cartoon character) or ideas (e.g., any combination of path, manner, ground components of a motion). In other words, a gesture is ‘recruited’ by speech as such a sign. An emblem, such as an “okay” sign, is not exceptional in this regard: it also has to be explicitly recruited by speech to mean a certain thing at some point prior to its use among the members of the community which has it as a conventional sign. It just does not have to co-occur with speech each time it is used. The reversed situation can also be found: a pointing gesture, for example, ‘indexically entails’ the referent of a demonstrative that. This is an example of speech and gesture mediating each other and thereby achieving coreferentiality.

The extent to which speech is explicit in recruiting a gesture varies. In the maximally explicit case, the speaker performatively names a gesture, as is often done by an origami instructor who explains various complicated foldings without using a piece of origami paper.[16][17] For example, one can say, “This is a square piece of paper,” while tracing a square in the air with an index finger. Although a gesture of this kind often synchronizes with the co-expressive speech,[14] it need not synchronize with any text, as when a cataphoric this is used to index it. The virtual ‘square piece of paper’ established by what one might call a baptismal ritual of naming with speech and gesture at the center of the stage,[3] then, serves as a frame of reference, even though it is invisible in the discourse that follows. For instance, one can refer to the upper right corner of the square as in “You wanna take the upper right corner of the square” with a pointing directed at the upper right corner of the invisible square in the air, which is now shared among the communicative participants. Note that the speech, together with the concurrent gesture, entails that the corner pointed to by the pointing gesture is ‘the upper right’ corner of the square as opposed to something else, while it presupposes the perspective from which the imaginary paper is contemplated by the relative directional term, namely the upper right.

People do not always explicitly name a gesture in this way, however. This is often the case with a cartoon story narrator.[13] For example, one can say “The cat goes straight into the building” with her loosely extended right hand sweeping horizontally from the right to the left where the gesture represents the path of motion taken by the cat. For a gesture of this kind to be interpreted as ‘referring’ to something without being explicitly recruited to do so, what becomes extremely important are the speech segments a gestural segment co-occurs with. Hence, the tendency for tight synchronization between the two signs in such cases is key.[13] Conversely, a GP can be inferred from several indexical relationships. One is the instrumentally attested synchronization relationship of the stroke of a gesture and the acoustic peaks (e.g., the peaks of the F0 and the intensity) of the concurrent speech,[13] and the co-expressivity that is implied by the speech-gesture synchrony. The other is the indexical relationship of such a synchronous speech-gesture combination with the preceding and the following discourse segments, namely the discursive context in which GPs emerge. As the discourse unfolds, some elements of it last throughout the discursive event while new elements
are added. In short, there are always invariants and variants throughout the discourse. A GP indexes both a significant departure from the immediately preceding context (= variants) and a continuation of some part of the context (= invariants). Thus, it indexes the elements in the immediate context as newsworthy. To borrow Vygotsky’s words, GPs are “psychological predicates (as opposed to grammatical predicates).”[14] The common elements of the gestures that appear in different points of the discourse constitute what is called a catchment structure, of which an example is consistent uses of a particular handshape (e.g., G-handshape in ASL alphabet as in Figure 1), a gestural position, etc. A catchment structure, thus, indexes some aspects of the entire discourse as coherently connected to each other, and it is a backdrop against which a new discourse element emerges as a growth point. GPs, then, are “the seeds for a model of realtime utterance generation and coherent text formation.”[14] The validity of this is attested by a study on reference maintenance in Japanese discourse.[18][19]

3.3 Information Packaging Hypothesis (IPH)

The Information Packaging Hypothesis (or IPH for short) shares many theoretical points with GP theory, upon which it is based.[11] It argues for the interaction between analytic thinking and imagistic thinking, each of which underlies speech and gesture, respectively. It contrasts with GP theory in the following three points, however.

(1) The IPH does not concede analytic thinking and imagistic thinking as one unit of analysis. Instead, the IPH treats them as analytically separate by default, although this argument is not meant to deny, but rather support the interaction between the two modes of thinking:

what generates a gesture is spatio-motoric thinking, which has a raison d’être independent of speaking. Consequently, it is expected that gestures and speech have a certain degree of independence from each other.[11]

(2) Since the IPH treats the two modes of thinking as separate, it needs to be equipped with interface mechanisms which enable the system to assess “the degree of match between the representations in the two modes of thinking”[11] and to translate “representations in one mode of thinking into those in the other insofar as representational resources in each mode of thinking (such as lexicon) allow.”[11] GP theory does not and need not have such a translation mechanism.

(3) Finally, the IPH argues that, unlike GP theory, what underlies a gesture is not representational, but actional.[11] That is, the imagistic thinking underlying gesture derives from instrumental action in the real environment, of which an example is to grasp a glass of water.[11] Not only that, the IPH argues that gesture is an action in the virtual environment. Let us discuss the points (1) and (2) in the rest of this section, leaving (3) in the next, as a similar argument to this is also made by LeBaron and Streeck.[20]

(1) The advantage of permitting the separation of analytic thinking from imagistic thinking, it is argued, is that it “allows spatio-motoric thinking to explore different organizations of information, in some cases without fully coordinating with analytic thinking.”[11] Because of the separation, the IPH can allegedly explain the mismatch in the semantic contents between the concurrent speech and gesture found in children in the transitional phase of acquiring the concept of, for example, Piaget’s conservation of quantity and arithmetic problem solving as reported by Goldin-Meadow et al.[4] As the details of the description are crucial in the later discussion, let me extensively quote the description of the phenomenon by the original authors, beginning with that of the task of conservation of liquid quantity:

When asked to explain their judgments on the conservation tasks, the children gestured spontaneously while speaking and often portrayed specific aspects of the conservation task in their gestures; for instance, in a task probing conservation of liquid quantity, children often used a C-shaped hand to indicate the width of the dish, or they produced a fist and arced the fist from the glass to the dish as though pouring a pitcher to indicate that the water had been poured from the glass into the dish. At times the information conveyed in gesture matched the information conveyed in the speech accompanying the gesture. For example, in the task probing liquid quantity conservation, one child focused on the height of the water in both speech (“there is
less water in the dish because the dish is short and the glass is tall") and in gesture (the child demarcated the heights of the two containers of water with his palm). This child had thus conveyed a single hypothesis—expressed in both speech and gesture—to explain his solutions on the task.

In contrast, some of the gestures produced by the children did not convey the same information as did the speech accompanying those gestures. For example, in the liquid quantity conservation task, one child focused on the height of the container in speech ("the dish is lower than the glass") but focused on the width of the container in gesture (the child produced a wide C hand near the dish and a narrower C near the glass). This child had conveyed two distinct hypotheses to explain her solutions on the task—one involving height (in speech) and another involving width (in gesture).[4]

(The emphasis is by the present author. The ASL C hand is shown in Figure 3 below.)

![Figure 3: ASL alphabet C on its side.](image)

As for the task of conservation of arithmetic problem solving, Perry et al.[21]
tested children between the ages of 9 and 10 years on their understanding of equivalence in addition problems (i.e., the understanding that one side of an equation represents the same quantity as the other side of the equation). Children were asked to solve six problems of the form $5 + 3 + 4 = \_ + 4$ and to explain each of their solutions. Most children gestured spontaneously while explaining their solutions, and their gestures typically conveyed specific procedures for solving the problem. As in the conservation study, the procedure conveyed in gesture often matched the procedure conveyed in the speech accompanying that gesture. For example, one child indicated that he had added all of the numbers in the problem to get the answer, both in speech ("I added 5 plus 3 plus 4 plus 4 equals 16") and in gesture (the child pointed at the 5, pointed at the 3, pointed at the left 4, pointed at the right 4, and then pointed at the blank); that is, the child conveyed a single procedure, expressed in speech and in gesture.

However, again as in the conservation study, the gestures produced by the children did not always convey the same procedure as the speech that accompanied that gesture. For example, one child, in speech, indicated that he had added the numbers on the left side of the equation to get the answer ("I added 5 plus 3 plus 4") but, in gesture, indicated that he had considered all of the numbers in the problem (he pointed at the 5, the 3, the left 4, the right 4, and then the blank). This child conveyed two procedures, one that involved adding the numbers up to the equal sign (in speech) and a second one that involved adding all of the numbers in the problem (in gesture).[4] (The emphasis is by the present author.)

The mismatch-like phenomenon is reported to be found in adults as well, when they are having difficulty with packaging their thinking-for-speaking.[11] GP theory, on the other hand, is taken by the IPH as incapable of explaining the mismatch due to the lack of separation between the two modes of thinking; i.e., a GP cannot contradict within itself. Commenting on this, McNeill himself once said that "a possible relationship between the IPH and GP model is to see them as points along a continuum of communicative and operational effectiveness,"[22] suggesting the possibility that "GPs break down temporarily during mismatch phases of development, whereupon the situation described by the IPH takes over."[22] It is not entirely clear, however, how and when the shift between the IPH mode and the GP mode occurs.

The present paper argues that GP theory by itself can explain the mismatch of speech and gesture, and that the IPH has its own problems. First and foremost, in all these discussions, the interpretation of the "mismatch" is made from the perspective of adults who understand the concept of the conservation. One may wonder, however, whether it is also a "mismatch" for the children who produced it. We can and presumably should assume that it actually is not a mismatch for the children at least at the very moment of speaking/gesturing. In fact, it is exactly...
what makes it possible to interpret these as “mismatch” gestures that is missing in the children in the transitional phase.

Secondly, as we have seen in the lengthy description of the phenomenon quoted above, many of the gestures under discussion here are pointing gestures at least as far as the arithmetic task is concerned. This suggests that there is not really a “semantic” contradiction between what speech means and what gesture means. This is because the semanticity of a pointing gesture is minimum such that the referent cannot be determined without making reference to the context of its use, including what the closest object is on the extended line of the pointing index finger, what is indexically entailed and/or implicated by the concurrent speech, etc. It is at this pragmatic implicational level where the seemingly “semantic” contradiction in fact takes place. As such, the “mismatch” is even cancelable for these pointing gestures. Once we realize this, we should also be able to notice that the other non-pointing (so called iconic) gestures, at least in part, also operate at the pragmatic implicational level and are cancelable, because they do not have any meaningfulness outside of the context of their use. For example, the wide or narrow C-handshape gesture indicating the width of the dish or the glass in the quantity conservation task is motivated by, and hence indexes, one of the dimensions of the objects, in terms of which the gesture is interpreted. In other words, at least as far as the particular experimental setting is concerned, the iconic gestures also have a built-in indexical (and implicational) quality, if it is not the only semiotic quality, when these gestures are in use. Reconsidered this way, the “mismatch” phenomenon is not really a challenge to GP theory, because the speech-gesture mismatch as well as the match is pragmatic rather than semantic in nature. In fact, a pragmatic discrepancy between analytic thinking and imagistic thinking is fundamental and even necessary in the dialectic processes that GP theory invokes, since they necessarily presuppose a thesis and an antithesis. Needless to say, it would be natural for GP theory as well that the children in the transitional mismatch phase—having produced and retro- as well as extro-spectively seen the “mismatch” between speech and gesture—are more likely than those in the pre-transitional phase to detect the mismatch, from which they can step up to the new level of understanding. In a sense, the children in the transitional phase can take advantage of much wider area of attention. The same explanation should in principle also hold for similar phenomena found in adults as reported by Kita.

The argument of the IPH that the mismatch phenomenon is only possible due to the alleged ‘default’ segregation between analytic thinking and imagistic thinking would rather challenge the IPH itself. According to the IPH, the two modes of thinking work interactively when there is a discrepancy (a kind of mismatch) between them (as in the case of the celebrated ‘swing scene’ narration by Japanese speakers discussed above), and that they work separately or without much, if any, interaction otherwise. In the latter case, information is transparently translated from one mode of thinking into another without much negotiation. In the former case, quite importantly, the interaction of the two modes of thinking does not neces-

<table>
<thead>
<tr>
<th>Table 1: Different paths for a child to take, given different conditions and different levels of comprehension of the task.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mismatch between analytic thinking and imagistic thinking</strong></td>
</tr>
<tr>
<td><strong>Comprehension phase</strong></td>
</tr>
<tr>
<td>(a) No mismatch</td>
</tr>
<tr>
<td><strong>Pre-transitional phase</strong></td>
</tr>
<tr>
<td>(b) No mismatch</td>
</tr>
<tr>
<td><strong>Transitional phase</strong></td>
</tr>
<tr>
<td>(c) Mismatch</td>
</tr>
<tr>
<td>(d) Mismatch</td>
</tr>
</tbody>
</table>
sarily guarantee that it results in a speech-gesture match. This point could be made much clearer if we consider the different paths taken by the production system of speech and gesture and their consequences, given different phases of comprehension, as shown in Table 1 above.

If there is no mismatch between analytic thinking and imagistic thinking, there is no need for an interaction to take place between the two modes of thinking unless the interface mechanism (mentioned above) falsely detects a mismatch. This is indeed the case with children in the comprehension phase as in (a) in Table 1 and those in the pre-transitional phase as in (b) in Table 1 of understanding the conservation of quantity or arithmetic task. The result is a speech-gesture match, and this does not cause any trouble. If there is in fact a mismatch and if the system does not detect it, the two kinds of information might actually be processed separately, resulting in a speech-gesture mismatch, as the IPH argues. But, then, this cannot explain the synchronous relationship between the speech and the gesture, on which the analysis of mismatch bases itself. (The speech and the gesture are merely unrelated otherwise!) This is the case with children in the transitional phase (c) in Table 1. Further, if there is a mismatch and if the interface mechanism detects it successfully, the system would have the two modes of thinking work interactively. This could result in a speech-gesture match if the 'negotiation' between the two modes of thinking is successful, or just result in a mismatch due to an unsuccessful negotiation, or else even stop the entire production process. This is the case with children in the transitional phase (d) in Table 1. What is crucial here is that this suggests that the default segregation of the two modes of thinking is not a prerequisite to explain the mismatch phenomenon. This is because a mismatch can in principle be produced even when there is an interaction between the two modes of thinking. In this sense, the IPH is no better than GP theory.

(2) Let me make a short comment on the idea of the interface mechanisms. For a system to be able to determine the degree of match between the two kinds of representations and to translate one mode of thinking into another, the system must be equipped with a common code that can express both analytic thinking and imagistic thinking. If there is such a common code in the system, however, there need not be separate modes of thinking in the first place. Additionally, it is not entirely clear where such a common code ontogenetically and/or phylogenetically originates from.

4 Gesture-as-Action Hypothesis

The Gesture-as-Action Hypothesis basically makes the following argument:

Gestures, in our view, originate neither in the speaker's mind nor in the process of speaking, even though speech and gesture are routinely coordinated. Rather gestures originate in the tactile contact that mindful human bodies have with the physical world.

A similar idea is shared by the Information Packaging Hypothesis:

Spatio-motoric thinking can also be applied to the virtual environment that is internally created as imagery. Representational gestures are actions in the virtual environment.

In other words, in the Growth Point theory, what underlies a gesture is representational, and in the Information Packaging Hypothesis it is actional (emphasis in original).

Because of these common elements, then, let us treat LeBaron and Streeck's and Kita's claims together here, and refer to them as the Gesture-as-Action Hypothesis. In the rest of the present section, we will discuss the alleged evidence for the hypothesis. One type of evidence comes from an observation of gestures employed in the instruction of the manipulation or design of material objects. According to the authors, the gestures first appear as instrumental actions, e.g., the actual manipulations of objects in the real environment. They evolve, however, into symbolic displays, and then into "conventionally" fixed gestures commonly shared within a group of people who participated in the session. The original instrumental action at this point has become a symbolic reference. Another type of evidence comes from the finding of "gesturing during silence or meaningless vocalization and a "semantic" mismatch between gesture and
speech during transition phases of conceptual development.\footnote{11}

The present paper considers these findings extremely important.\footnote{6} Yet, it argues against the interpretation of the phenomena made by the Gesture-as-Action Hypothesis. It is certainly not the case that there is no element of truth about the claim that gesture can bring in information that is at least difficult, if not impossible, to convey by means of speech alone. However, there are at least two reasons that cast doubts on the claim that gesture originates from instrumental action in the real environment. One reason comes from the neurologically informed finding of gestural performances of a patient (IW) who “as a young adult suffered an infection that caused the loss of all proprioceptive feedback and spatial position sense from the neck down,” but fortunately spared “all functions above the neck, including speech and cognition.”\footnote{24} The Gesture-as-Action Hypothesis claims that gesture is an action in the virtual environment which has its origin in instrumental action in the real environment where the gesturer interacts with the real objects. On this view, the patient without any proprioceptive feedback or spatial position sense from the neck down would have difficulty with gesture if his body is made invisible, since he has difficulty with instrumental action without visual feedback. It turned out, however, that the patient gestured perfectly normally (in terms of timing and co-expressivity with the concurrent speech) under the condition where he could not even see his body from the neck down.\footnote{24} GP theory maintains that “gesture is not instrumental action but is action onto which meaning is mapped.”\footnote{24} In other words, “gesture is an action that helps to create the narrative space that is shared in the communicative situation, and as such, it comes under the control of linguistic and communicative systems rather than the instrumental motor system.”\footnote{24} This naturally explains the perfectly normal gestural performances of the patient.

Another reason that also casts doubt on the Gesture-as-Action Hypothesis comes from LeBaron and Streeck’s own observation cited in Example 1 below.\footnote{20} Notice that all of the actions/gestures documented occurred in the discursive event, namely the instruction of the manipulation or design of material objects where the instructor ceaselessly talks about the materials and what to do with them. In fact, we can find that speech is crucially involved in the very beginning of the “symbol formation” and throughout this discursive event as a whole.\footnote{20} Observe the excerpt of the instructor’s speech and actions/gestures in Example 1:\footnote{20}

Example 1. An example of a gestural symbol formation in an instruction session.\footnote{7}

\textit{Instructor:}

1.1 There’s a couple of things you need for preparing Sheetrock.

\begin{tabular}{l}
((picks up scraper))
\end{tabular}

1.2 (— — — — )

((puts scraper in other hand, looks at it))

1.3 One of them will be the scraper of some sort.

\begin{tabular}{l}
((scrapes in mid-air with both hands; puts it down))
\end{tabular}

1.4 This is a uh— very heavy-duty scraper.

\begin{tabular}{l}
((picks up other scraper; holds it up))
\end{tabular}

Commenting on this, LeBaron and Streeck say:

The instrument is picked up from the table and turned into an exhibit when it is moved from one hand to the other while simultaneously receiving the instructor’s concentrated gaze (line 2, ...). The tool’s use is then demonstrated through a schematic motion in a virtual field of action—that is, the instructor ‘scrapes’, but performs in the air (line 3, ...).\footnote{20}

Notice, first of all, how densely the actions/gestures are surrounded by words in the example above. Note also that the utterance in 1.4, among others, counts as a speech act of naming.\footnote{25}\footnote{26} It is not in the form of what one might call the “explicit performative sentence,” but is rather somewhat indirect as a speech act of naming. Nonetheless, what is going on here in the “symbol formation” is naming a certain object which, in this case, is a scraper, and it shares essentially the same basic formula with the metalinguistic (i.e., metasemantic and metapragmatic) discourse that involves the copula, i.e., $A \textit{is B}$\footnote{27}\footnote{28}
This “baptismal” event of naming things, as it were, serves as a Putnamean stereotypy which is presupposedly indexed in the successive discourse by those who witnessed the “baptism,” forming an “indexical chain” of referential acts.[29] The reference can thereafter be accomplished by means of producing a token of the word and/or the gesture, which are iconic to the original ones (in terms of the sound image in the case of speech, and in terms of the handshape, the gestural position, the motion, and so on, in the case of gesture), so that the members of the community can recognize them despite the locality of the group (i.e., the participants of the session or the group of people who has engaged in the same activity). Gestures with the similar handshape and motion recursively appear, as they are reported, after the line 1.4 of the example with and without the real object, the without-case being an example of a gesture in the evolved form; i.e., the “symbolic gesture.”[20]

Note that the baptismal ritual is not always achieved with the explicit formula. Thus, the speaker in a cartoon narration often presents a gesture without verbally specifying what it represents in the form of “A is B” unlike the instruction session reported by LeBaron and Streeck.[20] In such a case, the indexicality, or the poetic function in Jakobson’s sense, foregrounds.[27] That is to say, the synchronization of a gesture with the meaningfully related speech segments and the indexing of a gesture with the eye gaze, to list just a few, would do the job. So, for example, a token of the ASL G-hand and the eye gaze indexing it, both of which are synchronized with the speech segment “the cat” in the cat goes up the drainpipe can be taken as representing the animal despite the absence of likeness between the sign and the animal and the absence of the baptismal ritual in the explicit formula of “A is B.” The same holds for the use of like this/that that is accompanied by a (pointing) gesture. Even the mismatching gesture does not suggest that “gestures and speech have a certain degree of independence from each other,”[41] considering the indexical relationship between the two as discussed above.

In sum, the Gesture-as-Action Hypothesis is not as convincing as it first appears to be. Speech and the discursive event as a whole play a crucial role in the genesis and evolution of a gesture in the speech/gesture community. The same holds true even when speech does not synchronize with the gesture, but merely precedes and/or follows it. This makes perfect sense if we think about the notion of indexicality which is not at all restricted to the synchronization, but covers the spatiotemporal contiguity in general, including the successiveness as well as simultaneity regarding the “temporal” contiguity, and any type of arrangement regarding the “spatial” contiguity.[3] GP theory maintains, I repeat once again, that “gesture is not instrumental action but is action onto which meaning is mapped.”[24] This, however, should not be taken as suggesting that gesture does not obey the dynamical laws or regularities that other instrumental actions do. The claim here is that gesture is mediated by language.

5 Conclusions
In this paper, we discussed several psycholinguistic
theories of speech-gesture coordination proposed in the recent years, and attempted to determine which theory is most promising and convincing. (See Table 2 for summary of the present discussion.) Although I will not repeat the reasons provided in the above discussions here, it appears that the Growth Point theory is more promising than others that have been discussed in this paper.

Regardless of which viewpoint one might want to take, all of the theories would be challenged by what one might call inter-personal coordination of speech and gesture as opposed to intra-personal speech-gesture coordination.[16][17][30][31][32] I will close the present paper, though, while leaving that issue to be discussed in a companion paper.[32]

Acknowledgments
The present paper, together with its companion paper,[32] has elaborated upon a short paper read at the 1st International Workshop on Interactive Graphic Communication (IGC 2000) held at Queen Mary and Westfield College, University of London, in London, U.K. in August 30 and 31, 2000.[30] I received insightful and useful comments from the participants, and they have made my thinking about the issues dealt with here much clearer. David McNeill at The University of Chicago read parts of earlier versions of the present paper and gave me invaluable comments. Mischa Park-Doob proofread the entire texts. In the early phase of writing up this paper, I received a William Harper Dissertation Fellowship from the University of Chicago (2000-2001), and in the final phase of the write-up, I received a research grant from the Japan Society for the Promotion of Science for a research project on speech-gesture coordination—the central theme in the present paper (Grant #13224095, 2001-2002). I would like to thank all of these individuals and organizations for their help, encouragement and understanding.

References
[16] Furuyama, N., “Gestural interaction between the instructor and the learner in origami instruction”,
A Review of Theories of Speech-Gesture Coordination

Unpublished MS thesis submitted to Dept. of Psychology, University of Chicago, 1996.


Endnotes

1 To be sure, Japanese has a transitive verb fur-u (to swing non-present) that encodes the arc trajectory of the patient’s swing movements.

2 McNeill and his colleagues observe that gestural performances can be different depending on what language one speaks, thus demonstrating that gestures can be influenced by different languages in different ways.[14]

3 Notice that the performativity of speech is at work[25], which is metasemantic[27] as well as metapragmatic in nature[23] That is, the demonstrative this explicitly indexes the concurrent gesture and the utterance as a whole explicitly spells out its meaning.

4 McNeill’s notion of material carriers of thinking is identified as “a more actional view on gesture in the Growth Point theory.”[11] As we will see below, however, McNeill does not take a standpoint that gesture originates in the instrumental action.[23] The Growth Point theory will be introduced in the next section.

5 Judging only from the contrast made in this quotation, the Information Packaging Hypothesis seems to argue that what underlies a gesture is not representational. However, the IPH also talks about a translation between
different “representations” in the same article.

6 As a matter of fact, the present author demonstrates that the learners in origami instruction frequently gesture without producing any speech by themselves.[16][17]

7 The convention here is the following: Square brackets “[”indicate the simultaneity of two events, either two or more utterances or an utterance and a physical action or gesture; Double brackets “()” mark the transcriber’s description of an action; Dashes in parentheses “(— — —)” indicate a pause and each dash represents approximately one-tenth of a second.[20]