Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Memory and Language 48 (2003) 16–32
Journal of Memory and Language
www.elsevier.com/locate/jml

What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking

Sotaro Kita[a,*] and Asli Özyürek[b]

[a] *Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, Netherlands*
[b] *Koç University, Istanbul, Turkey*

Received 1 May 2001; revision received 14 February 2002

Abstract

Gestures that spontaneously accompany speech convey information coordinated with the concurrent speech. There has been considerable theoretical disagreement about the process by which this informational coordination is achieved. Some theories predict that the information encoded in gesture is not influenced by how information is verbally expressed. However, others predict that gestures encode only what is encoded in speech. This paper investigates this issue by comparing informational coordination between speech and gesture across different languages. Narratives in Turkish, Japanese, and English were elicited using an animated cartoon as the stimulus. It was found that gestures used to express the same motion events were influenced simultaneously by (1) how features of motion events were expressed in each language, and (2) spatial information in the stimulus that was never verbalized. From this, it is concluded that gestures are generated from spatio-motoric processes that interact on-line with the speech production process. Through the interaction, spatio-motoric information to be expressed is packaged into chunks that are verbalizable within a processing unit for speech formulation. In addition, we propose a model of speech and gesture production as one of a class of frameworks that are compatible with the data.

Keywords: Semantic coordination; Cross-linguistic comparison; Speech production; Gesture production; Motion event

This paper investigates the cognitive process that underlies spontaneous co-speech gestures, especially its relationship to speech production. Theories of gesture production differ in how gestures are informationally related to the content of concurrent speech and at what level of the speech production process the content of gestures is determined. There are three hypotheses regarding these issues: The Free Imagery Hypothesis (de Ruiter, 1998, 2000; Krauss, Chen, & Chawla, 1996; Krauss, Chen, & Gottesman, 2000), the Lexical Semantic Hypothesis (Butterworth & Hadar, 1989; Schegloff, 1984), and the Interface Hypothesis. These hypotheses make different predictions as to how the content of gestures may differ cross-linguistically when speakers describe certain spatial events. This study aims to contrast these three hypotheses by comparing gestures that are produced by speakers of Japanese, Turkish, and English.

Some theories of gesture production maintain the Free Imagery Hypothesis. According to this hypothesis, gestures are generated from imagery in working memory

* Corresponding author.
E-mail addresses: kita@mpi.nl (S. Kita), aozyurek@ku.edu.tr (A. Özyürek).

0749-596X/02/$ - see front matter
PII: S0749-596X(02)00505-3

and their content is constructed on the basis of long-term memory of events or some other thought processes. More importantly, they are generated "prelinguistically," that is, independently from the representational potential of the language. Krauss et al. (1996, 2000), for example, suggest that gestures are generated from the spatial imagery in the working memory, which is activated at the moment of speaking. Unlike Krauss et al. (1996, 2000), de Ruiter proposes that representational gestures are generated by the process that also generates speech, namely the Conceptualizer (in the sense of Levelt, 1989), which produces a pre-verbal message to be fed into the linguistic formulation module. However, the models proposed by Krauss and his colleagues and by de Ruiter are similar in that gestures are generated before linguistic formulation processes take place. Consequently, the Free Imagery Hypothesis predicts that the information encoded in a gesture is not influenced by how the information could be verbally expressed.

In contrast, other theories maintain the Lexical Semantics Hypothesis, where gestures are generated from the semantics of lexical items in the accompanying speech. For example, Butterworth and Hadar (1989) claim that a lexical item generates iconic gestures through one or more of its semantic features that can be interpreted spatially. In other words, iconic gestures are generated from the result of the computational stage in speech production after the "selection of the lexical items in abstract form from a semantically organized lexicon" (Butterworth & Hadar, 1989, p. 172). The idea of certain lexical items being the source of iconic gestures was originally proposed by Schegloff (1984), who claims that "various aspects of the talk appear to be 'sources' for gestures affiliated with them" (Schegloff, 1984, p. 273). He further notes that the source is the "lexical components of the talk" (Schegloff, 1984, p.275). The prediction of the Lexical Semantic Hypothesis is that representational gestures do not encode what is not encoded in the concurrent speech.

The third view is the Interface Hypothesis, which we propose in this paper. According to this view, gestures originate from an interface representation between speaking and spatial thinking. The interface representation is the spatio-motoric representation (i.e., information about action and spatial information represented in terms of action) that is organized for the purpose of speaking. Thus, according to the Interface Hypothesis, gestures not only encode (non-linguistic) spatio-motoric properties of the referent, but also structure the information about the referent in the way that is relatively compatible with linguistic encoding possibilities. This hypothesis is based on the following view of speech production processes.

To speak, the information to be expressed has to be tailored for speaking. Namely, "thinking for speaking" (Slobin, 1987, 1996) is necessary. More specifically, the information to be expressed has to be organized so as to include the information necessary for obligatory morphological markings (Slobin, 1987), and to be made more compatible with the lexical and constructional resources of the language (Slobin, 1996). Furthermore, the information to be expressed has to be adapted to the linear nature of speech (Levelt, 1989) and the limited capacity of the speech production system. Rich and complicated information has to be organized into smaller packages so that each package has the appropriate informational complexity for verbalization within a processing unit for speech production. This unit corresponds to what can be processed within one processing cycle for the formulation of speech. Thus, the optimal informational organization for speech production for a given language is determined by interaction between representational resources of the language and processing requirements for the speech production system.

The necessity for organizing information for speaking becomes clear in light of cross-linguistic variation of how certain concepts are linguistically expressed. A certain concept may correspond to a readily accessible concise expression in one language but not in another. For example, in some languages, it is not straightforward to translate the English sentence, "Tarzan swung across the street," because they do not have an intransitive verb that has the equivalent meaning to the English verb, "to swing." A certain concept may be equally expressible in different languages, but with different levels of linguistic complexity. For example, expressing certain aspects of a motion event may require only one clause in one language but multiple clauses in another language (Talmy, 1985).

It has been argued that such linguistic differences indeed influence how spatio-motoric representations of the referent are prepared in the course of speech production, and they are visible in speech accompanying gestures (Kita, 1993, 2000a,b, 2002, in press; McNeill, 1992, 2000; McNeill & Duncan, 2000; Özyürek & Kita, 1999). This argument is based on the Growth Point Theory of utterance generation put forth by McNeill (1992), where the planning of utterances involves the interplay of imagistic thinking and linguistic thinking. The outcome of imagistic thinking manifests itself as gesture and the outcome of linguistic thinking manifests itself as co-expressive speech. It has also been argued that gestures are generated from a process by which spatio-motoric imagery is shaped into a form that is suitable for speaking (Alibali, Kita, & Young, 2000; Kita, 1993, 2000a,b, in press). In this view, gestures are involved in the process of packaging the spatio-motoric imagery into informational units suitable for speech production. The process of linguistically formulating ideas in speech has capacity limitations and there is an optimal linguistic unit for this process. We call such a unit a *processing unit* for speech production.

A processing unit can roughly be approximated by a clause (Bock, 1982; Garrett, 1982; Levelt, 1989). Thus, informational units suitable for speech formulation are what can be encoded in a clause in a given language.

This leads to the Interface Hypothesis for the representational characteristics of gesture. The Interface Hypothesis states that the spatio-motoric imagery underlying a gesture is shaped *simultaneously* by (1) how information is organized in the easily accessible linguistic expression that is concise enough to fit within a processing unit for speech production and (2) the spatio-motoric properties of the referent (which may or may not be verbally expressed). That is to say, the hypothesis predicts that a gesture is shaped by the formulation possibilities of the language (unlike the Free Imagery Hypothesis) and at the same time the gesture may encode the spatio-motoric information that is not expressed in the speech (unlike the Lexical Semantics Hypothesis).

Note that the Interface Hypothesis is distinct from a hybrid hypothesis, based on the Lexical Semantic Hypothesis and the Free Imagery Hypothesis; namely, some gestures are generated in the way suggested by the Lexical Semantic Hypothesis and others are generated in the way suggested by the Free Imagery Hypothesis (Hadar & Yadlin-Gedassy, 1994). The Interface Hypothesis predicts that for a given gesture one can observe the simultaneous influence of both the linguistic formulation possibilities and the spatio-motoric properties of the referent that are not verbalized in the accompanying speech.

Hadar and Butterworth (1997) propose a model of speech and gesture production, which also proposes interplay between imagistic and linguistic processes. However, their proposal differs crucially from ours in that the relevant linguistic unit for the interplay is a single word. In contrast, the relevant unit in our proposal is an informational unit that can be linguistically encoded within a processing unit for speech production, which is approximately a clause. [See also de Ruiter (1998, 2000) for other arguments for positing a unit larger than a word as the relevant unit.]

The goal of this paper is the following. First, we provide evidence from a cross-linguistic comparison of speech–gesture coordination that supports the Interface Hypothesis, but not the Free Imagery Hypothesis or the Lexical Semantics Hypothesis. Furthermore, we will also argue that the relevant unit for the linguistic effect on gestural representation is an informational unit that corresponds to a processing unit for speech production (approximately a clause). The results from this study constrain possible models of how processes of speech and gesture production are inter-related. In addition, we propose a model of speech and gesture production as one of a class of frameworks compatible with the data.

Present study

In this paper, the above predictions of different hypotheses are tested with a cross-linguistic comparison. The test ground is created by cases where languages package information differently for certain types of stimulus events. The Interface Hypothesis predicts that gestural expressions are simultaneously shaped by linguistic formulation possibilities and by the spatial properties of the events that may not be linguistically encoded in the accompanying speech. Specifically, the Interface Hypothesis predicts that the gestural expression of the events varies across languages in ways similar to the linguistic packaging of information about the events in respective languages.

The languages to be compared are American English, Turkish, and Japanese. We analyzed gestures that are produced in narratives elicited from the same stimulus. We focused on the gestural expression of two scenes, in which the three languages differ in how they package information.

In the first scene, due to the limitation of the linguistic expressive resources of Turkish and Japanese, which makes it difficult to verbalize a certain prominent spatial feature of the event (i.e., the arc trajectory of a motion). In contrast, this feature is easily encodable in English. Thus, as part of the conceptual planning for speaking, it is desirable for Turkish and Japanese speakers to generate a representation of the event without the feature that is difficult to verbalize. In contrast, English speakers can keep the prominent spatial feature as a part of the representation of the event. The Interface Hypothesis predicts that this cross-linguistic difference in preparation for speaking will be reflected in the gestural representation of the event. In other words, the feature that is difficult to verbalize is less likely to be gesturally represented by Japanese and Turkish speakers than by English speakers.

In the second scene, two simultaneous features of the event to be described are linguistically packaged more concisely in English than in Turkish and Japanese. Consequently, Turkish and Japanese speakers are more likely to spread the simultaneous features over two or more processing units for speech production, whereas English speakers are more likely to package the two features into one processing unit. The Interface Hypothesis predicts that in Japanese and Turkish, it is more likely that two separate gestures will be used to represent the two features, whereas in English the two features are more likely to be simultaneously encoded in one gesture.

Furthermore, the Interface Hypothesis predicts that the gestures that show the influence of the linguistic formulation possibilities will also regularly encode some spatial details that may not be verbally expressed in the accompanying speech. This is because gestures are

generated from imagistic representations of the referent events. When translocational motion is represented as imagery, certain features of the event, such as the direction of the motion, have to be specified regardless of their significance in the discourse. In the two scenes discussed above, whether the lateral motion was to the right or to the left is not consequential in the plot development and thus this information is not likely to be expressed in speech. However, when the motion is represented as imagery, its direction has to be specified. Thus, the gesture that is generated on the basis of the imagery should regularly encode the direction of the motion based on the visual experience of the stimulus.

The Free Imagery Hypothesis predicts that there is no cross-linguistic difference in the gestural content for both the first and second scenes that we just discussed, but that gestures regularly encode spatial details that may not be verbally expressed. The Lexical Semantics Hypothesis predicts that gestures reflect differences in linguistic encoding possibilities in the three languages, but that gestures do not regularly encode spatial details that are not verbalized.

To obtain a cross-linguistically comparable gesture corpus, narratives in American English, Japanese, and Turkish were collected using the same stimulus. The methodology basically follows that of McNeill (1992).

Method

Participants

Sixteen adult native speakers of American English, 18 adult native speakers of Turkish, and 17 adult native speakers of Japanese participated in the experiment.

Materials

The stimulus was an American animated cartoon, which was about 6 min long. The recurrent theme of the cartoon was a cat's (Sylvester) unsuccessful attempts to catch a bird (Tweetie). For a detailed description of the cartoon, see the appendix of McNeill (1992).

Procedure

Each participant was told that they were participating in a story telling experiment. She/he was instructed to remember the stimulus as well as possible so as to be able to tell a detailed story to a person who did not see the stimulus. Gesture was not mentioned in the instruction. The participant was shown the stimulus on a TV monitor, while the listener waited in another room. Immediately after watching the stimulus, the participant told the story to the listener. No specific instruction was given to the listener except that he/she should pay attention to the story and was allowed to ask questions. Each participant's narration was videotaped.

Effect of limitation in linguistic expressive resources on gestural representations

The first analysis is carried out to investigate how limitation in expressive resources of a given language affects gestural representation. The scene in the stimulus that is selected for the analysis is the Swing Scene. In the Swing Scene, a cat and a bird are across the street from one another in the windows of different high-rises. The cat's building is on the right side of the screen and the bird's building is on the left side of the screen. In an attempt to catch the bird, the cat swings across the street on a rope that we must imagine is attached somewhere in the air above the street. Fig. 1 is the schematic drawing of the event.

In Turkish and Japanese, there is no readily accessible expression that semantically encodes agentive change of location with an arc trajectory. There is no verb that corresponds to the English intransitive verb "to swing" as in "the cat swings across the street". There is no readily accessible paraphrase for it either. (It would be possible to use mathematical terms like "arc" to paraphrase English "swing" such as "fly, drawing an arc," but such a paraphrase would not be a readily accessible one.) Thus, this is not only a lexical gap, but it is also a more general limitation in the expressive resources of the two languages.

This cross-linguistic difference requires speakers of the three languages differ in their conceptual planning for speaking. Turkish and Japanese speakers have to construe the Swing Event in such a way that the trajectory shape is abstracted out, whereas English speakers' construal of the event can include the arc trajectory. The Interface Hypothesis proposes that the spatio-motoric representation of the event, which manifests itself as gesture, reflects the way the speakers of each language

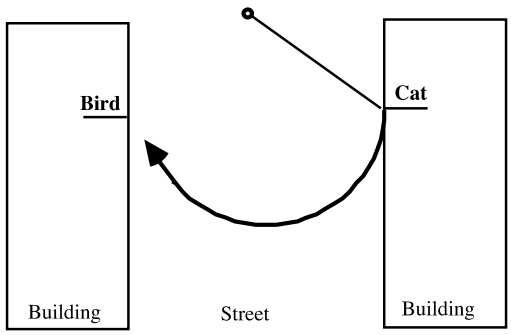

Fig. 1. The schematic representation of the Swing Event in the stimulus.

package the information about the event. Thus, it is predicted that Turkish and Japanese speakers are more likely to gesturally represent the event without the trajectory shape than American English speakers.

Furthermore, the Interface Hypothesis also predicts that the gestural representation of the event regularly reflects some aspects of the stimulus scene that are not expressed in the accompanying speech. It has been reported that the direction of the lateral movement (i.e., to the left or to the right) in the stimulus is regularly reproduced in the gesture, but rarely in the speech (McCullough, 1993). If the participant sees a movement in the stimulus that goes to the right on the video monitor, she/he is highly likely to gesturally represent the event as a movement to the right from the speaker's point of view. It is predicted that Turkish, Japanese, and American English speakers all regularly represent the lateral direction of the cat's change of location in their gestures, despite the fact that the content of these gestures is also shaped by the information packaging possibility of the respective languages.

Coding

The portion of the narratives in the three languages that referred to the change of location of the cat in the Swing Scene, henceforth the Swing Event, was analyzed. Gestures that expressed horizontal displacement were coded by two coders for the following two form features. First, it was coded whether the trajectory shape is "arc" or "straight". A gesture was coded as "arc" when its trajectory was downward concave (e.g. a semi-circle with the upward "opening," or any arc that is a part of such a semi-circle). A gesture was coded "straight" when it did not include downward concave trajectory. The second formal feature coded was the horizontal direction of the gesture: "left-biased" or "right-biased" or "purely away from the body."

Gestures by three randomly selected speakers from each language were used to check the inter-coder reliability. The nine speakers from the three language groups produced a total of 16 gesture tokens depicting the Swing Event. The two coders agreed on the arc–straight judgement on 94% of the tokens, and on the direction judgement on 87% of the tokens.

Results

Speech

All 16 American English speakers encoded the Swing Event in the speech. All but one used the word "swing" to describe the event. Fifteen (out of 17) Japanese speakers and 17 (out of 18) Turkish speakers encoded the Swing Event in the speech, but none of them lexically encoded the arc-shaped trajectory. Instead, they described the event with a change of location predicate that is trajectory-neutral. In Japanese, the verbs used in the description include "iku" (to go), "tobu" (to jump/fly), "shinobikomu" (to sneak in). In Turkish, the verbs used include "gidiyor" (to go), "ucuyor" (to fly), and "atliyor" (to jump).

With regard to the coding of the lateral direction of the swing event, none of the speakers of any of the languages used the words, "left" or "right."

Gesture

Trajectory shape encoding. Two English, one Turkish, and two Japanese speakers were excluded from this analysis because they either did not mention the target event or did not have a gesture with horizontal dislocation for the event.

The remaining participants were classified into three mutually exclusive categories according to their gestural behavior: those who used, in their description, arc gestures only, those who used both arc gestures and straight gestures, and those who used straight gestures only. Fig. 2 shows the percentage of the participants in the three languages who fell into the three categories. The proportions of the three categories of participants differed across the three languages (χ^2 test, $\chi = 12.167$, $DF = 2$, $p = .002$). The pattern of usage of arc and straight gestures was very similar between Turkish and Japanese speakers. More of the Turkish and Japanese speakers as a group used at least one straight gesture (i.e., the dark bar plus the gray bar in Fig. 2) than the English speakers (Fisher's exact test, one-tailed, $p < .001$).

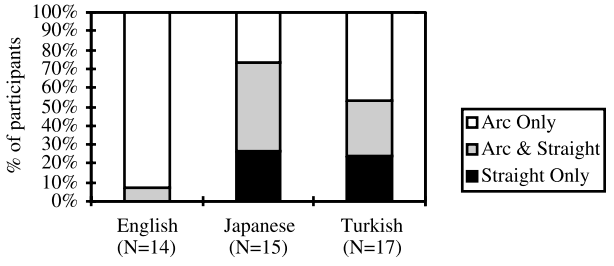

Fig. 2. Percentage of participants with the three patterns of usage of arc and straight gestures.

The gestural content varies cross-linguistically in a manner that parallels to how the three languages package information about the Swing Event in speech. This gestural variation across the languages is predicted by the Interface Hypothesis, but not by the Free Imagery Hypothesis.

Lateral direction of the movement. The direction of motion can be gesturally expressed from two different perspectives. One perspective is from the protagonist at his/her source location for the motion (event-internal perspective). In this perspective, the protagonist's body is mapped onto the speaker's body and the motion in the stimulus is expressed as a movement away from the body in gesture. Another perspective is from the viewer of the stimuli (event-external perspective). In event-external perspective, the viewer's body is mapped onto the speaker's body and the lateral motion in the stimulus, like the Swing Event, is expressed as a lateral movement in gesture. In our analysis, we focus on the gestures with event-external perspective because they allow us to test whether the gestural direction matches or contradicts the direction of motion in the stimulus.

It was found that the leftward motion in the stimulus (from the viewpoint of the viewer) was regularly reproduced in gesture, regardless of the trajectory shape. Tables 1 and 2 list the percentages of gesture tokens (aggregated over participants) that fell under the three categories of the horizontal direction coding.

The majority of the gesture tokens with event-external perspective encoded the lateral direction of the Swing Event as viewed by the participants (i.e., left-bias), and there were very few tokens that went the other way (i.e., right-bias). As we will see in the next section, when the target event was to the right (the opposite the direction of the Swing Event), the direction of gestures exhibited strong right bias. McCullough (1993) analyzed gestures elicited with the same stimulus with the same method as in this study and found also that the left–right directions of various stimulus events were consistently reflected in gesture directions. Thus, it can be concluded that regardless of the trajectory shapes and the language types, the Swing-Event gestures regularly encode the directional information in the Swing Event that is never verbalized. This is predicted by the Interface Hypothesis, but not by the Lexical Semantics Hypothesis, according to which gestures encode only what is encoded in the speech.

Discussion

Gestural expression of the Swing Event shows both systematic cross-linguistic variation and similarity, as predicted by the Interface Hypothesis. The cross-linguistic variation in the gestural representation of the Swing Event has the same pattern as the variation in the linguistic packaging of information about the event. In English, where there is a readily accessible linguistic means to package the change of location and the arc-shaped trajectory, speakers' gestures represent change of location with an arc-shaped trajectory. By contrast, in Turkish and Japanese, where readily accessible linguistic means cannot encode the arc trajectory, the majority of the speakers produced a change of location gesture without the arc-shaped trajectory. These findings demonstrate a linguistic effect on the gestural representation.

The Swing Event gestures, however, regularly encode spatial information that is not encoded in the speech. The lateral bias of the Swing Event gestures encodes the leftward movement in the stimulus, regardless of the encoding of the arc, in all three languages. The gestural representation reflects directional properties of the spatial information in the stimulus that is never linguistically encoded. The existence of arc gestures in Turkish and Japanese makes the same point. Furthermore, the

Table 1
The lateral bias of arc gesture tokens ("left" and "right" are from the viewpoint of the speaker)

Language	*N*	Event-external perspective		Event-internal perspective
		Left bias	Right bias	Purely away from the body
Turkish	20	85%	0%	15%
Japanese	23	74%	0%	26%
English	22	77%	0%	23%

Table 2
The lateral bias of straight gesture tokens ("left" and "right" are from the viewpoint of the speaker)

Language	*N*	Event-external perspective		Event-internal perspective
		Left bias	Right bias	Purely away from the body
Turkish	10	60%	0%	40%
Japanese	23	62%	8%	31%
English	4	75%	0%	25%

English verb "to swing" does not entail an arc movement on a vertical plane; for example, the word "swing" can also refer to an arc movement on a horizontal plane. However, the arc gestures in the English sample all represent an arc on a vertical plane, which is how the event happens in the stimulus (see Fig. 1). This is also an example of systematic coding of spatial information that is not in the speech.

The systematic encoding of the directional information in gesture provides a strong argument against the Lexical Semantic Hypothesis because the directionality is neither encoded in, nor inferable from the lexical items uttered by the speakers. Even the Turkish and Japanese straight gestures, which unequivocally demonstrate the linguistic effect on the gestural representation, clearly encode the directional information that is not expressed in the accompanying speech. Therefore, a gesture is *simultaneously* shaped both by readily accessible, concise linguistic packaging of relevant information and by the spatio-motoric properties of the referent that are never verbalized. This makes it difficult to maintain a hybrid theory between the Lexical Semantic Hypothesis and the Free Imagery Hypothesis, where some gestures are generated by the manner advocated by the Lexical Semantics Hypothesis and others are generated by the manner advocated by the Free Imagery Hypothesis.

Effect of different clausal packaging of spatial information on gestural representation

Another scene in the stimulus, where the three languages package information differently is the Rolling Scene. In this case, all three languages have readily accessible means to express the same aspects of an event. However, the linguistic package for the same information is tighter in English than in Turkish or Japanese due to the difference in the lexicalization pattern. The scene in question is the following: A cat, who has swallowed a bowling ball, has a big round stomach and bottom, and he rolls down the street into a bowling alley. (This movement was from the left to the right of the screen.) A few moments after he enters the bowling alley, there is a sound of pins being knocked down. The event in this scene, for which the three languages package information differently, is the one where the cat rolls down the street (henceforth the Rolling Event). The Rolling Event is the focus of analysis in this section.

Two components of the Rolling Event are lexicalized differently in English, on the one hand, and in Turkish and Japanese, on the other hand. The components are Manner, namely the rotation, and Trajectory, namely the continuous change of location of the moving entity. English typically encodes Manner in a verb and Trajectory in a preposition or a verb particle, whereas Turkish and Japanese typically encode both Manner and Trajectory in verbs (along the line of the linguistic typology proposed by Talmy (1985)). Thus, English can encode the event with a single clause, as in (1). By contrast, Turkish and Japanese use two clauses to encode the event, as indicated by the square brackets in (2a) and (2b). (Note, however, that Turkish has a more marked option to encode both Manner and Trajectory in one clause, an example of which will be given in the section "Speech".)

(1) He rolls down the hill.

(2)

a. *Japanese*

[korogat-te]	[saka-o	kudaru]
roll-Connective	slope-Accusative	descend: Present

"(s/he) descends the slope, as (s/he) rolls."

b. *Turkish*

[yuvarlan-arak]	[cadde-den	iniyor]
roll-Connective	street-Ablative	descend: Present

"(s/he) descends on the street, as (s/he) rolls."

The two components of the event are encoded in a tighter linguistic package in English than in Turkish and Japanese. Consequently, it is more likely that English speakers formulate both Manner and Trajectory within a processing unit for speech production.

The Interface Hypothesis predicts the information packaging in gesture to be similar to the information packaging in the accompanying speech. Namely, it is predicted that there is a tendency for Turkish and Japanese speakers to encode Trajectory and Manner in separate gestures, whereas English speakers put them together in one gesture (Özyürek & Kita, 1999). It is also predicted that speakers of all the languages preserve the non-linguistic structure of the event (the rightward direction of the Trajectory).

Coding

The part of the narratives in the three languages that refers to the change of location of the cat on the street (i.e., the Rolling Event) is analyzed. Gestures that depicted the Rolling Event were coded by two coders for the following two form features. First, gestures that accompany the Rolling Event description are categorized into three types: Trajectory Only, Manner Only, and Manner–Trajectory Conflating. A Manner Only gesture represents the circular nature of the rolling, and/or the repetitive aspect of rolling (e.g., a repetitive up and down movement of the hand), without representing change of location of the moving entity. A Trajectory Only gesture represents change of location without any Manner representation. In a Manner–Trajectory Conflating gesture, the representations of Trajectory and Manner are superimposed (e.g., a hand sweeping

horizontally, as it makes a small repetitive up and down movement). Trajectory Only gestures and Manner–Trajectory Conflating gestures are further coded for the horizontal direction of the gesture: "left bias" or "right bias" or "purely away from the body."

Gestures by three randomly selected speakers from each language were used to check the inter-coder reliability. The nine speakers from the three language groups produced a total of 23 gesture tokens depicting the Rolling Event. The two coders agreed on the judgement regarding the gestural content on the 100% of the tokens. For the 17 tokens that were judged to be Trajectory Only and Manner–Trajectory Conflating gestures, the two coders agreed on the judgement of the Trajectory direction on 100% of the tokens.

Results

Speech

Fifteen (out of 16) English speakers explicitly encoded the Rolling Event (i.e., change of location on the street) in the speech. All of them used one clause to encode Manner and Trajectory. The Manner verb, "to roll," was accompanied by a preposition or a verb particle such as "down," "along," and "across" (one speaker additionally produced an utterance without Trajectory encoded, "he is rolling").

Fourteen (out of 17) Japanese speakers explicitly encoded the Rolling Event. Thirteen speakers encoded both Manner and Trajectory, and all but one of them used two clauses to do so. One speaker used an ungrammatical expression, "michi-o korogat-te" (rolling the street), where a Manner verb is combined with a Trajectory-encoding postpositional phrase. In all other utterances, Trajectory encoding postpositional phrases, such as "on the street," "to the bowling alley," were syntactically associated with a Trajectory verb. The Manner verb was "korogaru" (to roll), which was often accompanied by a sound symbolic adverbial "gorogoro" (rolling continuously). The Trajectory verbs were "iku" (to go), "kudaru" (to descend), and "ochiru" (to fall).

Sixteen (out of 18) Turkish speakers explicitly encoded the Rolling Event. All speakers used separate clauses to do so except for one speaker who used an adverbial "yokus-asagi" (downhill) with a Manner verb, "kayiyor" (to slide). Otherwise, Trajectory encoding postpositional phrases, such as "on the street," "along the slope," "to the bowling alley," were syntactically associated with a Trajectory verb. The verbs that encoded Manner were "yuvarlaniyor" (to roll), "kayiyor" (to slide), "zipliyor" (to jump), and "sallaniyor" (to shake). A sound symbolic word, "dangir dungir" (repetitive sound made by a heavy object), was also used to encode Manner. The Trajectory verbs used were "gidiyor" (to go), "geciyor" (to cross), and "iniyor" (to descend).

To summarize, English speakers used one clause to encode both Manner and Trajectory in the Rolling Event. In contrast, there was an extremely strong tendency for Turkish and Japanese speakers to use separate clauses for Manner and Trajectory.

Gesture

Encoding of manner and trajectory. Two English, one Turkish, and three Japanese speakers were excluded from this analysis because they either did not mention the Rolling Event or they did not have any Manner Only, Trajectory Only, and Manner–Trajectory Conflating gestures for the Rolling Event. For each gesture type, for each language, we calculated the proportion of the participants who used the type of gesture in question at least once out of all the participants who gesturally represented the Rolling Event in one way or another. (Note that a given speaker could produce more than one type of gesture.)

Turkish and Japanese speakers patterned together in the usage of Trajectory Only and Manner Only gestures. They were different from English speakers in the way predicted by the Interface Hypothesis: Compared to English speakers, Turkish and Japanese speakers were more likely to have Manner Only and Trajectory Only gestures as part of their repertoire of gestural representations of the Rolling Event. As Fig. 3 shows, the proportions of participants that produced at least one Manner Only gesture was higher in Turkish and Japanese as a group than in English (Fisher's exact test, one-

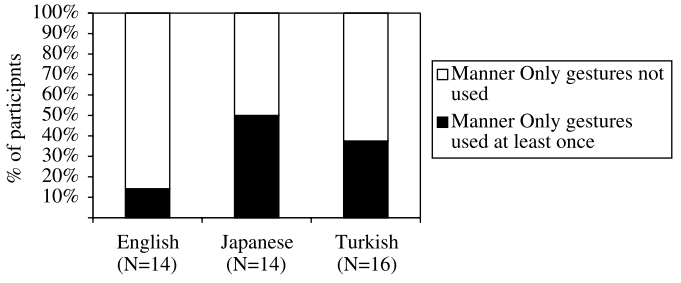

Fig. 3. Percentage of participants who used a Manner Only gesture at least once in their description of the Rolling Event.

tailed, $p = .045$). Similarly, as Fig. 4 shows, the proportions of participants that produced at least one Trajectory Only gesture was higher in Turkish and Japanese as a group than in English (Fisher's exact test, one-tailed, $p = .045$). This parallels the tendency that in the Turkish and Japanese speech Manner and Trajectory are more separated than in English.

In contrast, the speakers of the three languages were the same with respect to the likelihood of using a Manner–Trajectory Conflating gesture (Fig. 5). That is, the repertoire of gestural representations for the Rolling Event in all three languages were equally likely to include a Manner–Trajectory Conflating gesture. Note that a Manner–Trajectory Conflating gesture had the same structure as the Rolling Event in the stimulus, in that Manner and Trajectory were simultaneously realized.

Even though the three languages look the same in the analysis illustrated in Fig. 5, the three languages differ in whether other types of gestures are produced in addition to Manner–Trajectory Conflating gestures. In English, Manner–Trajectory Conflating was often the only type of gesture in the speaker's repertoire, whereas in Turkish and Japanese, the speakers who used a Manner–Trajectory Conflating gesture also used a Manner Only gesture and/or a Trajectory Only gesture. The dark part of the bars in Fig. 6 shows the proportion of participants who used Manner–Trajectory Conflating gestures alone

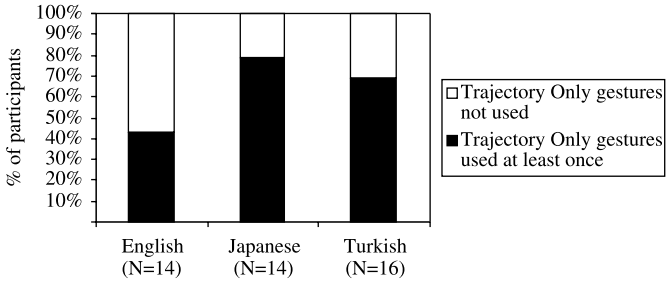

Fig. 4. Percentage of participants who used a Trajectory Only gesture at least once in their description of the Rolling Event.

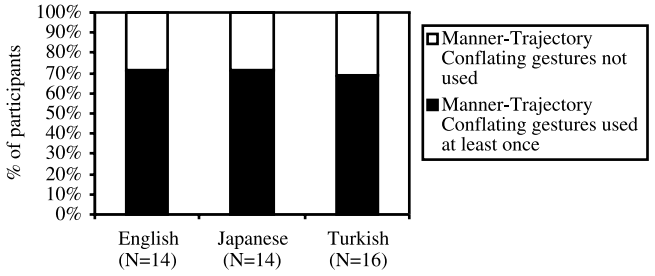

Fig. 5. Percentage of participants who used a Manner–Trajectory Conflating gesture at least once in their description of the Rolling Event.

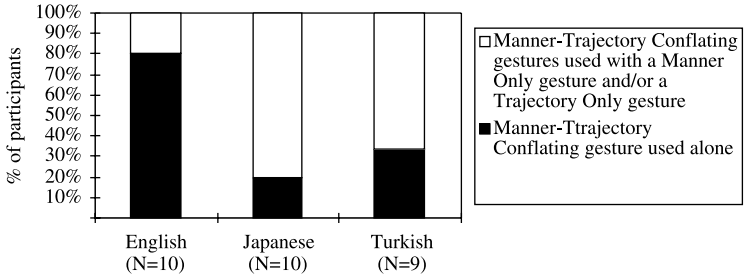

Fig. 6. Percentage of participants who used Manner–Trajectory Conflating gesture alone, and who used a Manner–Trajectory Conflating gesture in combination with Manner gesture and/or Trajectory Only gesture in their description of the Rolling Event.

in their description. (The "N" for a given language in Fig. 6 is the number of all the participants who used Manner–Trajectory Conflating gesture at least once). The proportions of participants that produced only Manner–Trajectory Conflating gestures were higher in English than in Turkish and Japanese as a group (Fisher's exact test, one-tailed, $p = .016$).

Even though speakers of the three languages were equally likely to have Manner–Trajectory Conflating gesture as part their repertoire, the status of Manner–Trajectory Conflating gesture in the description of the Rolling Event were not the same in the three languages. For Turkish and Japanese speakers, it was not sufficient to have a construal of the event that is similar to the non-linguistic structure of the Rolling Event. They had to further come up with informational chunks that were more compatible with their linguistic formulation possibilities, as can be seen in the additional use of Manner Only and Trajectory Only gestures.

Lateral direction of the movement. Trajectory Only and Manner–Trajectory Conflating gestures also regularly encoded the directional information in the stimulus event in the same way across the three languages, as predicted by the Interface Hypothesis. We again focus on the gestures with event-external perspective, which moved laterally, because they allow us to test our hypothesis. It was found that the rightward motion in the stimulus was regularly reproduced in both Trajectory Only gestures and Manner–Trajectory Conflating gestures, when they took the event-external perspective (Tables 3 and 4).

Though not directly relevant to the hypotheses to be tested in this paper, it is also interesting to note that Turkish speakers produced more gestures with the event-internal perspective, which moved away from the body, than Japanese and English speakers. A further investigation is necessary to determine what causes Turkish speakers to diverge from Japanese and English speakers in terms of the gestural perspectives for the Rolling Event.

The direction of Trajectory Only gestures is of special theoretical interest. As shown in Fig. 4, the usage of Trajectory Only gestures varied cross-linguistically in a way that made gestural representations of the Rolling Event similar to how the concurrent speech packaged information about the event. Yet, the same Trajectory Only gestures encoded the direction of the event that was never verbalized. In other words, gestures exhibited simultaneously the influence of linguistic packaging of information and the structure of the spatial representation of the event to be described.

Discussion

The gestural representation of Manner and Trajectory shows both cross-linguistic variation and similarity as predicted by the Interface Hypothesis. The repertoire of gestural representations for the Rolling Event differed in the way similar to linguistic encoding patterns in the three languages, namely, the compactness of the linguistic construction that expresses two simultaneous aspects of the event. More Turkish and Japanese speakers represented Manner and Trajectory in separate gestures than English speakers. In Manner–Trajectory Conflating gestures, Manner and Trajectory are simultaneously realized, which is also the case in the stimulus event to be described. Thus, this type of gesture is equally likely to be part of the repertoire of gestural representation of the Rolling Event in all three languages. However, Turkish and Japanese speakers who used a Manner–Trajectory Conflating gesture often combined it in the discourse with Manner Only gesture and/or Trajectory Only gesture.

Table 3
The lateral bias of Trajectory Only gesture tokens ("left" and "right" are from the viewpoint of the speaker)

Language	*N*	Event-external perspective		Event-internal perspective
		Left bias	Right bias	Purely away from the body
Turkish	13	0%	76%	24%
Japanese	10	10%	90%	0%
English	5	0%	100%	0%

Table 4
The lateral bias of Manner–Trajectory Conflating gesture tokens ("left" and "right" are from the viewpoint of the speaker)

Language	*N*	Event-external perspective		Event-internal perspective
		Left bias	Right bias	Purely away from the body
Turkish	9	11%	56%	33%
Japanese	11	9%	91%	0%
English	13	0%	92%	8%

The data on the direction coding in gesture provides evidence against the Lexical Semantic Hypothesis. A large majority of Trajectory Only gestures and Manner–Trajectory Conflating gestures regularly encode the lateral direction of the stimulus event, which is never verbally expressed. These data, especially regarding Trajectory Only gestures, indicate that gestures are shaped simultaneously by the spatial properties of the stimulus event and the linguistic encoding pattern. This argues against a hybrid hypothesis between the Lexical Semantics Hypothesis and the Free Imagery Hypothesis.

The above results also shed light on the issue of the linguistic unit relevant for the linguistic effect on the content of iconic gestures. Hadar and Butterworth (1997) proposed a model of speech and gesture production, which allows speech encoding possibilities to influence the gestural content. However, their model differs from the Interface Hypothesis in that the linguistic unit relevant for such an influence is a single word. Namely, in their models, the gestural content can be altered if there is no lexical item that encodes a certain set of semantic features. Such a model does not predict any difference in gestural content among Turkish, Japanese, and English speakers because the three languages do not differ in the availability of lexical items that encode Manner and Trajectory. The crucial difference among the three languages is that Japanese and Turkish require a more complex expression for Manner and Trajectory than English.

Consequently, it is likely that, in Japanese and Turkish, the speaker needs two processing units for speech production to express the two concepts, whereas English speaker needs only one processing unit. In the Interface Hypothesis, the linguistic unit relevant for linguistic effects on the gestural content is what can be verbalized within in a processing unit for speech production. Thus, the Interface Hypothesis predicts differences in the gestural expression of the Rolling Event by Turkish and Japanese speakers compared to English speakers.

In addition, some details of the data suggest that there is another factor that influences the content of gestures. The difference in frequency between Trajectory Only gestures and Manner Only gestures illustrates this point. More speakers use a Trajectory Only gesture than a Manner Only gesture in all three languages (Figs. 3 and 4). We suggest that this is due to the importance of Trajectory in the plot development. The change of location is necessary information leading the story to its dramatic ending, where the cat enters a bowling alley and then one hears the sound effect of pins being knocked down. This is consistent with McNeill's (1992) idea that discursively important information is more likely to be encoded in the gesture.

General discussion

Cross-linguistic variation of iconic gestures

The main finding of this study is the existence of cross-linguistic variation in iconic gestures. The language you speak affects the contents of iconic gestures. As the first approximation, iconic gestures for the same event are similar cross-linguistically. McNeill (1992) compared iconic gestures produced by speakers of Georgian, Swahili, Mandarin Chinese and English, who all described the same stimulus cartoon. He notes, "A remarkable thing about iconics is their high degree of cross-linguistic similarity. Given the same content, very similar gestures appear and accompany linguistic segments of an equivalent type, in spite of major lexical and grammatical differences between the languages. This resemblance suggests that the gesture emerges at a level where utterances in different languages have a common starting point—thought, memory, and imagery." (McNeill, 1992, pp. 221–222). However, his own current work (McNeill, 2000; McNeill & Duncan, 2000) and other work (Müller, 1998) show that iconic gestures can vary cross-linguistically. This paper, more specifically, demonstrates that gesture represents a spatial event in a way similar to how speech expresses the same event, but at the same time gesture includes spatial details that may not be expressed in the concurrent speech.

We have argued that the separation of Manner and Trajectory in Turkish and Japanese gestures is due to the fact that it is difficult to verbalize the two pieces of information within a single processing unit for speech production. Note that this explanation is not solely based on structural and lexical properties of the two languages. And, languages that are structurally and lexically similar to Turkish and Japanese, for example Spanish, may be different in terms of what information can fit into one processing unit. If so, gestures in these languages should exhibit different packaging of information from Japanese and Turkish.

This may account for a possible difference between Turkish and Japanese, on the one hand, and Spanish, on the other hand. All of the three languages typically need two verbs to express both Manner and Trajectory, unlike English (Talmy, 1985). However, there are some reports in the literature that suggest that Spanish speakers may typically conflate Manner and Trajectory in their gestures, as the English speakers in our study did. McNeill and Duncan (McNeill, 2000; McNeill & Duncan, 2000) suggest that speakers of Spanish may commonly conflate Manner and Trajectory in gesture though no quantitative data are reported in this regard. Senghas, Özyürek, and Kita (in press) report a similar finding though the sample size is small (four participants). The data on Spanish in the literature are not yet conclusive, but if Spanish is indeed different from

Turkish and Japanese, then we suggest the following explanation for the difference. It is possible that, compared to Turkish and Japanese, it is easier in Spanish to linguistically encode Manner and Trajectory within one processing unit for speech production. Spanish allows Manner verbs to be combined with a directional expression with a preposition "hasta" (up to), such as "rodó hasta la pista de bolos" ((s/he) rolled up to (until) the bowling alley), and Spanish speakers may use such a combination more widely than Japanese and Turkish speakers. In the Turkish and Japanese description of the Rolling Event, directional postpositional phrases were never used with a Manner verb alone. In addition, Spanish speakers may produce a Manner verb and a Trajectory verb in adjacent positions (or very close to each other) within one intonational phrase, similarly to a sequence of a Manner verb and a Trajectory particle or preposition in English. That is, Spanish speakers may have access to a construction in which a Manner verb and a Trajectory verb are tightly linked. In Turkish, a Manner verb and a Path verb are commonly separated by a phrase such as "the street", as in the examples in (2b). In Japanese, Manner information is often expressed in a sound symbolic word [see Kita (1997, 2001) for further information about this class of words in Japanese], which is typically intonationally separated from the Trajectory expression.

We have argued that the cross-linguistic differences in gestural representation of motion events emerge in the course of on-line planning for speech production. However, there is a possible alternative explanation along the lines of the linguistic relativity hypothesis as proposed by researchers such as Whorf (1939/1956), Lucy (1992), Pederson (1995), Pederson et al. (1998), and Levinson (1997, in press). That is, it is possible that Japanese and Turkish speakers' memory of the stimulus was shaped by the language they speak, and the representations in their memory are different from those of English speakers. For example, Japanese and Turkish speakers might have remembered the Rolling Event as consisting of two separate events, a rolling event and a change of location event. This alternative explanation, however, is not tenable. In the Turkish and Japanese sentences in (2), in fact, the morpheme that connects the clauses (i.e., "-te" and "-arak") explicitly indicate that Manner and Path are simultaneous aspects of a single event. This suggests that, just like English speakers, Turkish and Japanese speakers remembered Manner and Trajectory as two simultaneous aspects of a single event.

Furthermore, the alternative explanation is also unlikely for the cross-linguistic differences in gestural representations for the Swing Event. This is because the lexical gap in Turkish and Japanese seems to be accidental rather than systematic. For example, Japanese lacks an intransitive agentive verb of swinging, as mentioned above, but Japanese has a transitive verb of swinging ("furu") and an intransitive non-agentive verb of swinging ("fureru"), as in "a pendulum swings". It is not the case that Japanese, in general, avoids expressing an arc trajectory of a movement. If we assume that linguistic relativity of spatial memory arises from repeated exposure to a pattern of informational organization imposed by a language in the course of development, then it is implausible that the accidental gap in the Japanese and Turkish lexicons structures Japanese and Turkish speakers' memory in such a way that it filters out the arc trajectory of a movement. The lexicon of a given language is full of idiosyncrasy (in fact, a standard definition of lexicon is the depository of idiosyncratic information (e.g., Chomsky, 1965)). We argue that it is more plausible that adjustment of one's thought to the vast idiosyncrasy of the lexicon is performed on-line at the moment of speaking.

We have argued that the language specificity of gestural representation of motion events cannot be explained by language specificity of the memory of the events. However, our results may still have implications for linguistic relativity of thought. We have demonstrated that "thinking for speaking" postulated by Slobin (1987, 1996) is at work in the processing of non-linguistic spatial representation (see also McNeill, 2000). If language-specific spatial representation is repeatedly generated for speaking, then it can become part of habitual non-linguistic thought about space, that is, the default way of thinking about space even outside the context of speaking. At least, the current results that language can shape non-linguistic spatial representation in thinking-for-speaking opens the door to the possibility of language shaping thinking-in-general under certain circumstances.

A model of speech and gesture production

We have argued that the data presented in this paper support the Interface Hypothesis, but they are not compatible with the Free Imagery Hypothesis and the Lexical Semantic Hypothesis. This conclusion does not single out a particular model of speech–gesture production, but it constrains the type of possible models. We propose the following model as one of a class of theoretical frameworks compatible with the data.

The primary goal of the model is to specify how the content of a representational gesture is determined (and thus, phenomena concerning synchronization between speech and gesture are outside the scope of this model). The main characteristics of the model are graphically represented in Fig. 7. This model builds upon Levelt's (1989) model of speech production with some modifications and it incorporates ideas from Kita (2000a) and Özyürek (2002). Other models of speech and gesture production in the literature, such as those by de Ruiter

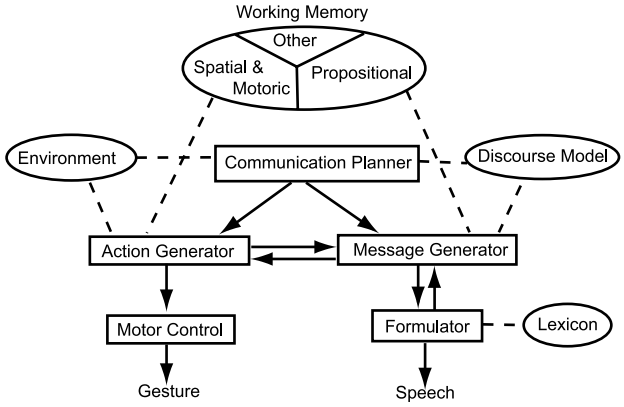

Fig. 7. Proposed model of speech and gesture production.

(1998, 2000) and Krauss et al. (2000), have also built upon Levelt's model.

Levelt's (1989) model of speech production makes a fundamental distinction between the planning process at the conceptual level ("Conceptualizer") and the speech formulation process ("Formulator"). The Conceptualizer transforms communicative intention into a propositional representation, called a "pre-verbal message," which is fed into the Formulator. The Formulator retrieves lexical items on the basis of conceptual specifications of the pre-verbal message and specifies the syntactic, morphological, and phonological make-up of an utterance.

In our model, Levelt's Conceptualizer is split into two halves. The first is the Communication Planner, which generates "communicative intention" and fulfils equivalent functions to Levelt's (1989) "macro-planning" (i.e., rough decision on information to be expressed, rough ordering of parts of the information for expression, and selection of appropriate speech acts). In addition, it determines which modalities of expression should be involved [incorporation of the modality selection process into the Conceptualizer is first proposed by de Ruiter (1998, 2000)]. The second half is the Message Generator, which fulfils functions equivalent to Levelt's (1989) "micro-planning" (i.e., formulating a proposition to be verbally formulated while taking into account both the communicative goal of an utterance and the discourse context).

The main characteristics of our model are the following:

1. The Communication Planner decides what modalities of expression should be involved, though it does not necessary determine exactly what information is to be expressed in each modality.
2. The content of a gesture is determined by
   (a) "communicative intention", generated in the Communication Planner,
   (b) action schemata selected on the basis of features of imagined or real space,
   (c) on-line feedback from the Formulator via the Message Generator. These three factors jointly determine gestural content and none of the factors alone fully specifies gestural content. In other words, gestural content is not fully specified in mechanisms dedicated to communication, such as Levelt's Conceptualizer, but rather in a more general mechanism that generates actions (Action Generator).
3. There is on-line bi-directional information exchange between the Message Generator and the Action Generator, and between the Formulator and the Message Generator. This allows gestural content to be shaped on-line by linguistic formulation possibilities.

The Communication Planner generates "communicative intentions" that grossly specify what needs to be communicated when. To take the example of the cartoon retelling task used in our study, a communicative intention might look like, "My global goal is to tell the story about the animated cartoon. Next, I want to describe the Swing Event, in which the cat tries to get to where the bird is in a particular way. I want to use both speech and gesture modalities for this purpose." This rough specification of the content to be expressed is sent to the Action Generator and the Message Generator. The Action Generator accesses the relevant part of the memory about the stimulus animation. Spatial imagery of the event, which is now active in working memory, includes both the arc trajectory and the directionality of the movement (i.e., to the left). We assume that the speaker's communicative intention does not include the directionality because none of our participants verbally expressed it even though in all three languages it would have been straightforward to do so. However, the directionality comes "for free" in the process of activating the spatial imagery of the event (to imagine a

translocational motion, one needs directionality, and the directionality can be obtained from the visual experience of the cartoon). Thus, the communicative intention determines the gestural content, but not fully.

The Communication Planner has access to the discourse model so as to take into account what has been communicated so far, and to project how the discourse should develop to achieve the overall goal of the discourse. Thus, the Communication Planner may give more prominence to certain information because of the goal of the discourse. For example, in the description of the Rolling Even in our study, Trajectory Only gestures are more common than Manner Only gestures across the three languages. We have argued that this is because change of location is essential information for the plot development and thus it is more like to be expressed.

The Action Generator is a general mechanism for generating a plan for action in real or imagined space [equivalent to "spatio-motoric thinking" in Kita (2000a)]. When an action is induced and guided by some features of space (e.g., grasping of an object), the action, in effect, selects those spatial features from a complex array of spatial information. Thus, generating such actions amounts to the parsing of space. This process is partly guided, for example, by what Gibson (1986) calls "affordances," structures that enable and induce certain action schemata in space. When an action is induced and guided by another action (e.g., mimicking an action by a protagonist in the cartoon stimulus), the newly generated action selects specific parts of the referred-to action. Thus, according to our model, gestures are generated from a general mechanism of action generation, which can be used in both purely communicative and practical purposes.

Since the Action Generator is a general process for generating actions, it has some degree of autonomy from the Message Generator as to which information to select from the environment or working memory. This leads to the issue of the interplay between the Action Generator and the Message Generator. The two Generators can independently initiate informational organization. Thus, there is no fixed order in which these processes operate (e.g., it is not necessarily the case that an image is first generated and then its content is passed onto the Message Generator). The two processes constantly exchange information and the exchange involves transformations between the two informational formats. A spatio-motoric representation, which is produced by the Action Generator, is transformed into a propositional format and passed onto the Message Generator. The Message Generator generates a proposition to be formulated in speech ("message"), which is transformed into a spatio-motoric format and passed onto the Action Generator. When the same communicative intention is given to the two Generators, the contents generated by these processes tend to converge through the exchange of information. (It is also possible for the Communication Planner to explicitly divide labor between the two modalities, for example, when gesture iconically demonstrates and speech indexes the gesture with an expression such as "like this". In this case, two coordinated but different goals are sent to the two Generators.)

The Message Generator, in addition, interacts on-line with Formulater. The message, generated by the Message Generator, is sent to the Formulator. If the proposition is not readily verbalizable within a processing unit, then the Message Generator receives direct feedback from the Formulator. In the case of Japanese and Turkish speakers describing the Swing Event, the Action and the Message Generators jointly explore and organize information about the event to specify exactly what information to express. During this process, the feedback from the Formulator to the Message Generator indicates that the trajectory shape is not readily verbalizable. This leads the Message Generator to take up the possibility of dropping the trajectory shape information. This new possibility is, in turn, translated into a spatial representation and passed onto the Action Generator. The Action Generator, the Message Generator, and Formulator keep exchanging information until equilibrium is reached, at which point formulation of speech starts and a spatio-motoric representation is sent to the Motor Control for execution of the movement (Kita, 2000a). This spatio-motoric representation in the Action Generator that is influenced by linguistic encoding possibilities is what we call the "interface representation" between speaking and the spatial thinking that makes use of action planning processes.

Note that the convergence of contents in the Action Generator and the Message Generator, on the basis of feedback from the Formulator, usually happens internally without overt vocalization or body movements [the process of convergence is, however, occasionally externalized, see Kita (2000a) for examples]. A certain level of convergence between the spatio-motoric representation and the message is required for initiating externalization of gesture and speech. The threshold, however, varies from moment to moment (Kita, 2000a). Such fluctuation can be seen in the Japanese and Turkish descriptions of the Swing Event. Japanese and Turkish speakers sometimes produce an arc gesture, which matches less well to the content of the concurrent speech, and they sometimes produce a straight gesture that matches better with the speech content.

Finally, as suggested by de Ruiter (1998, 2000), the Action Generator has access to the environment. It adjusts the shape of gestural representation according to the interactional and physical features of the environment. [See Özyürek (1997, 2000, 2002) for the effect of interactional features. The effect of physical features can be seen in an iconic gesture that traces the shape of an object in front of the speaker.] The Communication

Planner also uses information from the environment, such as the visibility of gestures from the addressee, which partly determines whether or not gestures are produced (e.g., Alibali, Heath, & Myers, 2001).

Relationship between the proposed model and other theories in the literature

Our model incorporates one of the key insights of the Growth Point Theory of gesture and speech production, originally proposed by McNeill (1992) and further elaborated by McNeill and Duncan (2000). That is, plans for co-expressive gesture and speech are shaped by dialectic between linguistic expressions and spatio-motoric representations, in which the two qualitatively different representations are adjusted with respect to each other and co-evolve.

According to our model, gestures are generated from a general mechanism of action generation (Action Generator), which can be used in both purely communicative and practical purposes. [Streeck (1996) and Müller (1998) maintain a related view that representational gestures have their origin in practical action.] This contrasts the view that gesture is generated by a mechanism that is dedicated solely for communication (de Ruiter, 2000; McNeill, 1992; McNeill & Duncan, 2000). Since the Action Generator in our model is a general mechanism for generating actions, it can select information with some degree of autonomy from the Message Generator. The autonomy of information selection allows content discrepancy between speech and gesture that seem to systematically occur when the speaker presumably cannot decide what exactly to say and use gestures to explore the possibilities (Alibali et al., 2000; Church & Goldin-Meadow, 1986; Goldin-Meadow, Alibali, & Church, 1993; Kita, 2000a, 2002).

Our model differs from other models in the literature with respect to the role of communicative intention. In de Ruiter's (2000) model, the gestural content is fully specified within the Leveltian Conceptualizer based on communicative intention. In contrast, Krauss et al. (2000) propose that communicative intention does not play any role in determining the gestural content in most gestures. We propose that communicative intention only roughly specifies the domain of information to be expressed, and the actual spatial and motoric information picked up by the Action Generator may include information, such as directionality of motion in our study, that was not part of the communicative intention.

Our model differs from Levelt's (1989) model of speech production and de Ruiter's (2000) model of speech–gesture production in that there is direct feedback from the Formulator to the conceptual planning level of speaking. We argue that the direct feedback is necessary to account for the fact that the informational content of Swing Event gestures is influenced by an idiosyncratic gap in the Japanese and Turkish lexicons. More generally, a direct feedback is necessary to adapt to the vast amount of idiosyncrasies in the lexicon, which may be as numerous as the number of all lexical items.

Concluding remarks

This paper provides some data that constrain the model of how gesture production and speech production processes are inter-related. More specifically, we have demonstrated that the content of representational gesture are shaped *simultaneously* by (1) how information is organized in the easily accessible linguistic expression that is concise enough to fit within a processing unit for speech production, and (2) the spatio-motoric properties of the referent, which may not be expressed in speech. On the basis of this finding, we have concluded that gestures are generated from the interface representation between speaking and spatio-motoric processes. In the interface representation, spatial and motoric information about the referent is packaged into chunks that are readily verbalizable within a processing unit for speech production. In addition, we have proposed a model of speech and gesture production as one of a class of frameworks compatible with the data.

Acknowledgments

This paper developed out of a chapter in Kita (1993), which discuss preliminary results on descriptions of the Swing Event in Japanese and English. Preliminary results on descriptions of the Rolling Event in Japanese and English were presented at the workshop "Gesture cross-linguistically" in Albuquerque, USA in 1995. The results from the comparison of English, Japanese, and Turkish with respect to the description of the Swing and the Rolling Events were presented at the 15th Japanese Cognitive Science Society Meeting in 1998 and at the 21st Annual Conference of the Cognitive Science Society in 1999. We acknowledge comments we received from the participants from these conferences. We also benefited from the feedback at different stages of this project from the members of the Gesture Project at Max Planck Institute for Psycholinguistics, especially Jan Peter de Ruiter. We benefited very much from the comments by Robert Krauss, Karen Emmorey, Alissa Melinger, and an anonymous reviewer on our manuscript. We thank David McNeill for the valuable comments on Kita (1993). We acknowledge the fact that the data from six American English speakers were kindly provided by David McNeill. The rest of the American English data and part of the Japanese data were collected with the resources of David McNeill's lab at the University of Chicago. The rest of the Japanese data and the Turkish

data were collected with the resources of Max Planck Institute for Psycholinguistics.

References

Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language, 44*, 169–188.

Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes, 15*, 593–613.

Bock, K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review, 89*, 1–47.

Butterworth, B., & Hadar, U. (1989). Gesture, speech, and computational stages: A reply to McNeill. *Psychological Review, 96*, 168–174.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

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

de Ruiter, J. P. (1998). *Gesture and speech production*. Ph.D. Dissertation, Nijmegen University.

de Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture* (pp. 284–311). Cambridge: Cambridge University Press.

Garrett, M. F. (1982). Production of speech: Observations from normal and pathological language use. In A. W. Ellis (Ed.), *Normality and pathology in cognitive functions* (pp. 19–76). London: Academic Press.

Gibson, J. J. (1986). *The ecological approach to visual perception*. Hillsdale, NJ: Lawrence Erlbaum.

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

Hadar, U., & Butterworth, B. (1997). Iconic gesture, imagery and word retrieval in speech. *Semiotica, 115*(1/2), 147–172.

Hadar, U., & Yadlin-Gedassy, S. (1994). Conceptual and lexical aspects of gesture: Evidence from aphasia. *Journal of neurolinguistics, 8*, 57–65.

Kita, S. (1993). *Language and thought interface: A study of spontaneous gestures and Japanese mimetics*. Unpublished doctoral dissertation, University of Chicago.

Kita, S. (1997). Two-dimensional semantic analysis of Japanese mimetics. *Linguistics, 35*, 379–415.

Kita, S. (2000a). How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture* (pp. 162–185). Cambridge: Cambridge University Press.

Kita, S. (2000b). Hito wa naze jesuchaa o suru noka [why do people gesture?]. *Ninchikagaku [Cognitive Studies], 7*, 9–21.

Kita, S. (2001). Semantic schism and interpretive integration in Japanese sentences with a mimetic: A reply to Tsujimura. *Linguistics, 39*, 419–436.

Kita, S. (in press). Interplay of gaze, hand, torso orientation and word in pointing. In S. Kita (Ed.), *Pointing: Where language, cognition, and culture meet*. Mahwah, NJ: Lawrence Erlbaum.

Kita, S. (2002). *Jesuchaa: Kangaeru karada* [Gesture: The body that thinks]. Tokyo: Kaneko Shobo.

Krauss, R. M., Chen, Y., & Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us?. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 28, pp. 389–450). Tampa: Academic Press.

Krauss, R. M., Chen, Y., & Gottesman, R. F. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), *Language and gesture* (pp. 261–283). Cambridge: Cambridge University Press.

Levelt, W. J. M. (1989). *Speaking*. Cambridge, MA: MIT Press.

Levinson, S. C. (1997). Language and cognition: The cognitive consequences of spatial description in Guugu Yimithirr. *Journal of Linguistic Anthropology, 7*, 98–131.

Levinson, S. C. (in press). *Space in language and cognition: Exploration in cognitive diversity*. Cambridge: Cambridge University Press.

Lucy, J. (1992). *Grammatical categories and cognition: A case study of the linguistic relativity hypothesis*. Cambridge: Cambridge University Press.

McCullough, K. E. (1993). Spatial information and cohesion in the gesticulation of English and Chinese speakers. A paper presented at the annual meeting of American Psychological Society, Chicago.

McNeill, D. (1992). *Hand and mind*. Chicago: University of Chicago Press.

McNeill, D. (2000). Analogic/analytic representations and cross-linguistic differences in thinking for speaking. *Cognitive Linguistics, 11*, 43–60.

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

Müller, C. (1998). *Redebegleitende Gesten : Kulturgeschichte-Theorie-Sprachvergleich*. Berlin: Spitz.

Özyürek, A. (1997). *Frames of reference and the effect of shared space on gesture and speech*. Unpublished doctoral dissertation, University of Chicago.

Özyürek, A. (2000). The influence of addressee location on spatial language and representational gestures of direction. In D. McNeill (Ed.), *Language and gesture* (pp. 141–161). Cambridge: Cambridge University Press.

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

Özyürek, A., & Kita, S. (1999). Expressing manner and path in English and Turkish: Differences in speech, gesture, and conceptualization. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the twenty first annual conference of the Cognitive Science Society* (pp. 507–512). Mahwah, NJ: Lawrence Erlbaum.

Pederson, E. (1995). Language as context, language as means: Spatial cognition and habitual language use. *Cognitive Linguistics, 6*, 3–62.

Pederson, E., Danziger, E., Wilkins, D., Levinson, S., Kita, S., & Senft, G. (1998). Semantic typology and spatial conceptualization. *Language, 74*, 557–589.

Schegloff, E. A. (1984). On some gestures' relation to speech. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social*

action: Studies in conversational analysis. Cambridge: Cambridge University Press.

Senghas, A., Özyürek, A., Kita, S. (in press). Encoding motion events in an emerging sign language: From Nicaraguan gestures to Nicaraguan signs. In A. Baker, B. van denBogaerde, O. Crasborn (Eds.), *Cross-linguistic perspectives in sign language research: Selected papers from TISLR 2000*. Hamburg: Signum Press.

Slobin, D. I. (1987). Thinking for speaking. In J. Aske, N. Beery, L. Michaelis, & H. Filip (Eds.), *Proceedings of the 13th annual meeting of the Berkeley Linguistic Society meeting* (pp. 435–445).

Slobin, D. I. (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.

Streeck, J. (1996). How to do things with things: Objects trouvés and symbolization. *Human Studies, 19*, 365–384.

Talmy, L. (1985). Semantics and syntax of motion. In T. Shopen (Ed.), *Language typology and syntactic description, Vol. 3, Grammatical categories and the lexicon* (pp. 57–149). Cambridge: Cambridge University Press.

Whorf, B. (1956). The relation of habitual thought and behavior to language. In language, thought, and reality. In J. B. Carroll (Ed.), *Language, thought, and reality: Selected writings of Benjamin Lee Whorf* (pp. 207–219). Cambridge: MIT Press (Original published in 1939).