

8

Trading Spaces: How Humans and Humanoids use Speech and Gesture to Give Directions

Justine Cassell, Stefan Kopp, Paul A. Tepper, Kim Ferriman, Kristina Striegnitz

Abstract Humans frequently accompany direction-giving with gestures. These gestures have been shown to have the same underlying conceptual structure as diagrams and direction-giving language, but the puzzle is how they communicate given that their form is not codified, and may in fact differ from one person to the next. Based on results from a study on language and gesture in direction-giving, we propose a framework to analyze such gestural images into semantic units (*image description features*), and to link these units to morphological features (hand shape, trajectory, etc.). This feature-based framework in turn allows us to implement an integrated microplanner for multimodal descriptions that derives the form of both natural language and gesture directly from communicative goals. In this way we have been able to realize an embodied conversational agent that can perform appropriate speech and novel gestures in direction-giving conversation with real humans.

8.1 Introduction

When describing a scene, or otherwise conveying information about objects and actions in space, humans make frequent use of gestures that not only supplement, but also complement the information conveyed in language. Just as people may draw maps or diagrams to illustrate a complex spatial layout, they may also use their hands to represent spatial information. For example, when asked how to find a particular address, it is common to see a direction-giver depicting significant landmarks with the hands—the fork where one road joins another, the shape of remarkable buildings, or their spatial relationship to one another.

Figure 8.1 shows just such an example of spontaneous gesture in direction giving. Here, the speaker has just said “If you were to go south”, then, while making this two-handed gesture, he says “there’s a church”. The gesture imparts visual information to the description, the shape of the church (left hand) and its location relative to the curve of a road (represented by the right arm). This meaning is instrumental to the understanding of the scene, and the listener took the gesture as intrinsic to the description.



Figure 8.1 Coverbal gesture on “There’s a church.”

But how do these gesture-speech combinations communicate? Unlike language, gesture does not display a lexicon of stable form-meaning pairings. And in virtual humans that are programmed to give directions to humans, how do we generate these same kinds of direction-giving units of speech and gesture? To date, research in this area has only scratched the surface of this problem, since previous ECAs draw upon a finite “gestonary” or lexicon of predefined gestures. While this approach can allow for the automatic generation of utterance, complemented by coordinated gestures, it provides us with neither an explanatory model of the behavior, nor the aspect of gesture that makes it so valuable—the communicative power and flexibility of our ever-present, visual, spatial modality.

In this paper, we start with a discussion of the role of the two generative systems of speech and gesture for giving directions. We will show that there are well-known, deep and fundamental differences between the kinds of information expressed by the two modalities, i.e. between the semantics of natural language and the meanings of gestures, and that these differences must be dealt with in order to understand and model the behavior. As posited by other researchers working on spatial language, we agree that this distinction necessitates the addition of an additional level of meaning for spatial and visual representations, beyond the two-level models of form and meaning seen in language. We report on an empirical study on spontaneous gesture in direction giving, whose results provide evidence for the existence of patterns in the way humans compose their representational gestures out of morphological features, encoding meaningful geometric and spatial properties themselves. It also suggests a qualitative, feature-based framework for describing these properties at the newly introduced level, that we call the *image description feature* level (IDF). We finally show how this approach allows us to model the multimodal generation problem by extending existing techniques for natural language generation. It lends itself to smooth integration of gesture generation into a larger system for microplanning of language and gesture, wherein linguistic meaning and structure can be coordinated with gesture meaning and structure at various levels. This approach allows us to capture the differences between the two systems, while simultaneously providing a means to model gesture and language as two intertwined facets of a single communicative system.

8.2 Words and Gestures for Giving Directions

Spatial descriptions including speech and gesture figure in descriptions of motion (Cassell & Prevost, 1996), in descriptions of houses (Cassell, Stone, & Yan, 2000), in descriptions of object shape (Sowa & Wachsmuth, 2003), in descriptions of routes (Tversky and Lee, 1998, 1999), and in descriptions of assembly procedures (Daniel, Heiser, & Tversky, 2003; J. Heiser, Phan, Agrawala, Tversky, & Hanrahan, 2004; J. Heiser & Tversky, 2003; J. L. Heiser, Tversky, Agrawala, & Hanrahan, 2003; Lozano & Tversky, submitted). The kinds of gestures implicated in these descriptions tend to be *iconic gestures*, where the form of the gesture bears a resemblance to what is represented by the gesture, and *deictic gestures*, that point out or indicate a path. In the remainder of this article, we concentrate primarily on iconic gestures, which can be an intrinsic part of spatial description.

Route directions are a particular kind of spatial description that is designed to assist a traveler in finding a way from point A to point B in an unknown environment. The description is typically organized into a set of route segments that connect important points, and a set of actions -- reorientation, progression, or positioning -- one of which is taken at the end of each segment. In order to ensure that the listener will be able to follow the segments, and to accomplish the right action, the speaker refers to significant *landmarks* (Denis, 1997). Landmarks are chosen for mention based on perceptual and conceptual salience (Conklin & McDonald, 1982), informative value for the actions to be executed, as well as visibility, pertinence, distinctiveness, and permanence (Couclelis, 1996). The quality of a route description is rated higher the more it consists of iterative steps of progression, pointing out landmarks, and reorienting the traveler (Denis, 1997).

Across all kinds of spatial description, researchers have found evidence that communicative behaviors portray a single underlying conceptual representation. For example, Cassell & Prevost (1996) in an investigation of manner-of-motion verbs and gesture in describing motion events, found ample evidence that the same concept may, in different situations, result in different realizations at the level of lexical items and paired gestures (e.g. “walked” vs. “went”+gesture). This suggests that communicative content may be conceived of in terms of semantic components that can be distributed across the modalities. In this study, roughly 50% of the semantic components of the described events were observed to be encoded redundantly in gestures *and* speech, while the other 50% were expressed non-redundantly either by speech or by gesture.

Similarly, an experiment on house descriptions (Cassell, Stone, & Yan, 2000) demonstrated that properties like shape, location, relative position, and path could be discerned in both speech and gesture. The particular distribution of properties, however, appeared to depend not only on the nature of the object described but also on the discourse context. For example, while the location of an object was redundantly conveyed by both speech and gesture,

properties such as contrast between two objects, their shape, location, relative position, and path through space most frequently occurred only in gesture while the existence of the objects was conveyed in co-occurring speech (e.g. “there was a porch” +gesture describing the curved shape of the porch).

When describing routes through environments too large to be taken in at a single glance, speakers adopt either a *survey* or a *route* perspective, or a mixture of both (H. A. Taylor & Tversky, 1992; Holly A. Taylor & Tversky, 1996). These two kinds of perspective on direction-giving result in two different patterns of gestures. When giving direction in a *route* perspective, speakers take their listeners on an imaginary tour of the environment, describing the locations of landmarks with respect to the traveler’s changing position, in terms of left, right, front, and back (“you walk straight ahead”). In a *survey* perspective, they adopt a bird’s eye viewpoint, and locate landmarks with respect to one another, in terms of an extrinsic reference frame, typically, north-south-east-west (e.g., “the house is south of the bridge”). Specifically, gestures during a route description tend to be in the plane in front of the body whereas gestures during a survey description are on a table-top or blackboard plane (Emmorey, Tversky, & Taylor, 2001).

In all of these cases, however, iconic gestures do not communicate independent of speech, for their meaning depends on the linguistic context in which they are produced. And listeners are unable to remember the form of gestures that they have seen in conversation (Krauss, Morrel-Samuels, & Colasante, 1991), although they do attend to the information conveyed in gesture, and integrate it into their understanding of what was said (Cassell, McNeill, & McCullough, 1999). How then, do we understand gesture?

8.2.1 *Arbitrary and Iconic Signs*

Despite their variety and complexity, when the findings described above have been applied to embodied conversational agents, the ECAs have followed the gestionary approach, i.e. they have been limited to a finite set of gestures, each of which conveys a predefined meaning. This is equivalent to a lexicon, wherein each word conveys a predefined meaning. The first step to move beyond this approach, both in understanding human behavior, and in computational models of virtual humans, is to realize that there are fundamental differences between the way iconic gestures and words convey meaning—differences that have already formed the basis of the study of signs by Saussure (Saussure, 1985) and Peirce (Peirce, 1955). Words are *arbitrarily* linked to the concepts they represent. They can be used to convey meaning because they are conventionalized symbols (“signifiers”), agreed upon by members of a linguistic community. Conversely, iconic gestures communicate through *iconicity*; that is, in virtue of their resemblance to the information they depict. These two “semiotic vehicles” also bear a markedly different relation to the context in which they are produced and interpreted. No matter what the context or the particular lexical semantics, relative to a gesture, a word always has a limited number of specific meanings. In contrast, an iconic gesture is *underspecified* (or indeterminate) from the point of view of the observer. That is, an iconic gesture has a potentially countless number of interpretations, or images that it could depict (Poesio, 1996) and is almost impossible to interpret outside of the context of the language it co-occurs with. For example, even limiting it to depictions of the concrete, the gesture shown in Figure 8.1 can be used to illustrate anything from the vertical movement of an object, to the shape of a tower, to the relative location of two objects, to a reenactment of a character performing some action. Clearly, it does not make sense to say that a gesture—observed as a stand-alone element separate from the language it occurs with—has semantics in the same way as language does when interpreted within linguistic context.

Even if an iconic gesture by itself does not uniquely identify an entity or action in the world, it always depicts (or specifies) features of an image through some visual or spatial resemblance. This is why we call it underspecified, and it is precisely this underspecification that is missing from most two-level models of form and meaning¹. Rather, to account for how iconic gestures are able to express meaning, we must have accounts of both *how images are mentally linked to entities* (the referents) and *how gestures can depict images*. Therefore, we conclude that to provide a way to link gestures to their referents, a third, intermediate level of abstraction and representation is required that accounts for a context-independent level of visuo-spatial meaning. While its application to gesture seems to be novel, the idea of such a representation is not new. It has further been advocated that this representation is multimodal or modality-independent (amodal), i.e., it underlies the processing of spatial information in different modalities, including speech and gesture. For example, Landau & Jackendoff (1993) discuss a spatial representation as “a level of mental representation devoted to encoding the geometric properties of objects in the world and the

¹ Although see (Poesio, 2005) for an approach to natural language processing based on underspecification.

spatial relationships among them” (p. 217). It is not exclusively visual or haptic or aural, but spatial, and we believe it to be drawn upon by language, gesture and the motor system more generally.

8.2.2 *Systematicity from Iconicity*

If iconic gestures are indeed communicative, people must be able to recover and interpret their meaning, and there must be a process by which we encode and, likewise, decode information in gesture. A reliable system for this requires some systematicity in the way gesture is used to depict, and the evidence from previous literature in several domains indeed suggests patterns in the form and function of iconic gestures with respect to expressing spatial information and communicating meaning more generally. Sowa & Wachsmuth (2003) report that one can find consistencies in the ways the fingers are used to trace a shape and that both palms may be held facing each other to illustrate an object’s extent. Unlike language, in gesture multiple form features may be combined to express multiple spatial aspects (e.g., extent and shape) simultaneously. Emmorey et al. (2001) observed that depictions of complex spatial structures are broken down into features that are then built up again by successive gestures. The fact that a single spatial structure is referred to across gestures (for example, a winding road) is signaled by spatial coherence; that is, the gestures employ the same viewpoint, size scale, and frame of reference, as indicated by a constancy of hand shape, trajectory and position in space. Sometimes, the frame of reference (e.g. relative to the winding road) is explicitly anchored in gesture space by one hand, and then held throughout while the other hand describes additional landmarks at appropriate relative locations. McNeill & Levy (1982) found positive and negative correlations for the association of distinct “kinesic” features in gesture, like fingers curled, palm down, or motion upwards, with semantic features of the motion verbs the gestures co-occurred with. For example, verbs with a horizontal meaning feature tended to co-occur with gestures with a sideways movement, but almost never with downward motion.

Originating in these results, our hypothesis is that there are prevalent patterns in the ways the hands and arms are used to create iconic, gestural images of the salient, visual aspects of objects/events, and that such patterns may account for the ways human speakers derive novel gestures for objects they are describing for the first time. However, we believe that the generativity that human gesture displays suggests that such patterning or commonality pertains not to the level of gestures as a whole, but to subparts—*features* of shape, spatial properties, or spatial relationships that are associated with more primitive *form features* of gesture morphology, like hand shapes, orientations, locations, movements in space, or combinations thereof. Consequently, we hypothesize that the intermediate level of meaning, which explicates the imagistic content of an iconic gesture, consists of separable, qualitative features describing the meaningful geometric and spatial properties of entities. We call these descriptors *image description features* (henceforth, IDFs).

Landau & Jackendoff (1991), as well as several others in this area, posit a range of geometric entities and spatial relations that seem to exist in such a mental spatial representation, based on analytical studies of linguistic data (Herskovits, 1986b; Talmy, 2000). They show that semantics of linguistic structures, e.g. prepositions or named objects, can be described in terms of these entities and spatial relations. For example, Landau & Jackendoff (1991) point out that words like *road* and *lake*, or similarly *wall* and *ceiling*, are often used to denote two-dimensional planes, or surfaces with negligible thickness. Many named objects are also imparted with intrinsic frames of reference, or directed axes, e.g. a front and back, top and bottom, left and right. Symmetric objects may also have sides, or ends. Words like *tall* or *wide* may be modified by dimensional terms which act with respect to the *primary axis* of an object. This axis would be dominant among an object’s intrinsic axes, e.g. a *tall person* or *building* modifies the vertical axis. These abstract spatial and visual features form the basis for the kinds of IDFs we propose.

We do not assume that human mental representations are necessarily feature-based, but assume that a feature-based approach will be adequate for describing these spatial and imagistic aspects. Such qualitative approaches have been successfully employed in research on spatial reasoning systems in the artificial intelligence literature (Forbus, 1983). Alternatively, the mental representations could be modeled using a more quantitative approach, e.g. using 3D graphics, allowing a more fine-grained, analog representation of the features discussed. Our approach has two advantages over this alternative: First, using quantitative features means that at some point, analog models must be annotated or linked to logical or semantic representations of what they depict. This is similar to the arbitrary way in which words are linked to their meanings in language. Second, our approach uses the same symbols at every level of the representation, so that expression of and computation on natural language semantics, gesture meaning and knowledge representation can all be carried out in a single, underlying representation. These advantages will be exploited in our computational model, as described further below.

Now we can define IDFs as links between gestures and the images they depict. They are features which can be used to describe the visual and spatial features of both a gesture’s morphology and the entities to which a gesture

can refer. So just as certain objects can be abstracted away from into classes based on visual and spatial features, certain hand shapes can be grouped into iconic categories, describing the image that they can carry. For example, just as *lakes*, *roads* and *walls* fall into the abstract class of *surfaces* or *planes*, likewise, certain hand shapes can be thought of as flat (2D), while others seem to have volume (3D). Our hypothesis is that these iconic gestures are composed of sets of one or more morphological features that convey sets of one or more image description features. We further conjecture that each of these mappings from IDFs onto form features can be found in different gestures depicting different, but visually similar things. Note, however, that this does not exclude the possibility of different morphological features being used to depict the same IDFs. That is, we are not assuming a one-to-one mapping here. We would have evidence for this hypothesis, if we could show that similar morphological features are generally used to depict similar visual or spatial properties.

Importantly, we are not claiming that there is such a thing as a “gesticon” or lexicon of gestures that consistently refer to objects or actions in the world. On the contrary, we are suggesting that *features* of gestures – handshapes, particular kinds of trajectories through space, palm orientations – refer to *features* of referents in the world – flatness in the horizontal plane, small roundness. It is this level of granularity in our hypothesis that allows us to explain how gestures can communicate, without having standards of form or consistent form-meaning pairings.

To illustrate our hypothesis, let us return to the utterance example in Figure 8.1. The subject’s right hand is held in place from the previous utterance and represents the curve in a road, anchoring the frame of reference. In the left-hand gesture, we find three form features: the flat hand shape with slightly bent fingers, the vertical linear trajectory, and the hand location relative to the right-hand gesture. If we suppose that each form feature corresponds to one or more IDFs in virtue of the resemblance of the former to the latter, we can analyze the gesture as follows: the relatively flat hand shape resembles a flat shape; or in more descriptive spatial terms, a two-dimensional, planar shape in a certain orientation. The vertical, linear trajectory shape corresponds to a feature that marks a vertical extent. Finally, the gesture location corresponds to a spatial location in relation to the frame of reference. All three IDFs in combination define an upright plane with a significant vertical extent, in a particular orientation and location. However, this content, which is inherent to the iconic gesture, does not suffice for a successful interpretation. Only when the gesture is placed in linguistic context, does the set of possible interpretations of the IDFs become so constrained as to make it unique. In our example, we infer from the indefinite noun phrase “a church” that the IDFs represent spatial information about the referent of the expression, namely a church. Linking the underspecified, imagistic features to this specific referent makes a successful interpretation possible, and we arrive at what McNeill (1992) deems the *global-synthetic* property of gesture, namely, that “the meanings of the parts of the gesture are determined by the whole (*global*), and different meaning segments are synthesized into a single gesture (*synthetic*)” (p. 41). The synthetic aspect is captured in some detail by the IDF approach; the global meaning of the gesture implies that the depicted upright plane becomes the wall of the church, viewed relative to the location of the road, and the vertical trajectory emphasizes the salient, vertical dimension, now corresponding to the height of the wall. Overall, we infer that the communicative intention of the speaker was to introduce a church, which has a tall, upright wall, and which is located near the curve of the road.

8.3 Relationship between Form and Meaning of Iconic Gestures in Direction Giving

To date, no literature describes in detail the interaction between iconic gestures and spatial language in route directions. Thus, to test our hypothesis, we collected video- and audiotapes of 28 dyads (more than five hours of dialogue) engaging in direction-giving. In each dyad, one person explained, without any external aids such as maps, a route from point A to point B on the Northwestern University campus to another person, who was unfamiliar with the campus. As mentioned above, speakers refer in such route descriptions to a constrained set of entities, including actions like progression (continuing along a path) and reorientation (turning), and objects like landmarks, their parts, shapes and spatial configurations (Denis, 1997). The Northwestern campus is ideal for this task as it provides numerous examples of objects (buildings, gates, bridges, etc.) that can serve as landmarks, while at the same time necessitating extensive and detailed instructions due to its size and complexity. This direction giving task thus demanded the speaker to communicate complex spatial and visual information only by means of the natural modalities. We expected that direction givers would make frequent and spontaneous use of coverbal iconic gestures (gestures that co-occur with language) to create representations of the spatial and visual information about landmarks and actions they needed to describe, and we were correct in this expectation. No subject neglected gesture in his or her description-giving.

In order to examine whether the IDF level of analysis is accurate, we *independently* (a) coded gestures into the features of their morphology, (b) coded referents in the world into their semantic features, and then (c) examined if

correlations existed between the two sets of features. If we find correlations between the two sets of features, then we have found evidence for IDFs, or the systematic use of certain visually similar classes of gestures to depict visually similar classes of referents.

Since all the directions were given around Northwestern campus, we were able to trace each route through the campus. Maps and photographs of the campus provided an independent source of information about the context of the utterances and gestures. This information allowed us to determine what seemed to be the “referent” for each phrase—e.g. the specific landmark, or aspects or subparts of it—and to determine its visual, spatial or geometric properties. To narrow down the scope of our exploratory study, we focused on gestures that seem to depict aspects of the shape of concrete objects, i.e. landmarks, parts of landmarks, streets and paths², as opposed to abstract entities like actions.

8.3.1 Method

Subjects

28 undergraduates, 11 men and 17 women, all native speakers of English, participated individually as direction givers in partial fulfillment of a course requirement. Three undergraduates unfamiliar with Northwestern University campus participated as direction followers; in the other 25 dyads, project personnel acted as naive subjects receiving the directions.

Materials

A list of ten routes was written, each of which consisted of five locations on campus to be visited in the order of listing. The first segment of each route, from the starting point to the first waypoint, was identical. The starting point was always the building in which the experiment took place. In addition, a schematic scale map of Northwestern University campus was printed.

Procedure

After arriving at the experiment room, the subject who was familiar with campus (the direction giver) was given the route list, and was asked to check every route she felt comfortable to give directions for. The subject was provided with the campus map in case she needed to look up names of waypoints or locations (see Fig. 8.2).

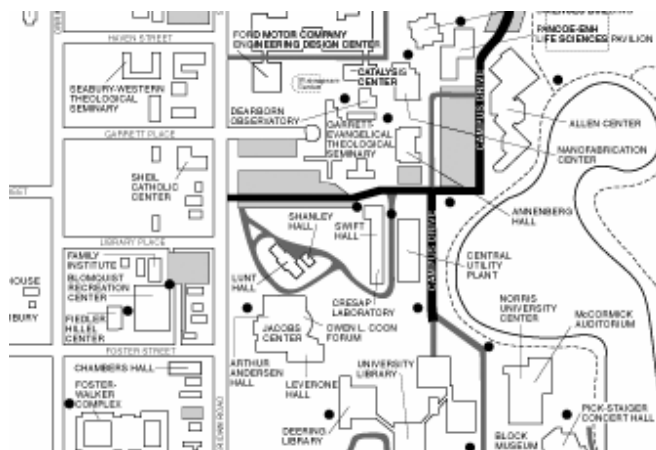


Figure 8.2 Example of campus map given to direction-giver for preparation

One of the checked routes was randomly chosen and assigned to the subject for description. In order to guarantee comparable conditions, the subject was instructed to familiarize herself with that particular route by then walking it herself. After the direction giver returned, she was seated face-to-face with the second subject (the direction follower) in a quiet room. Both subjects were instructed to make sure that the direction follower, who was—or pretended to be—totally unfamiliar with campus, understood the directions, and they were informed that the follower would have to find the route on her own, right after they concluded the session. Audiotapes and videotapes

² By path here we mean real paved or dirt paths around campus, not abstract paths of motion or trajectory.



Figure 8.3 Sample of the video data gathered (the arrow is inserted here to indicate the direction of movement).

	61.92	62	63
Words	you	face	the parking lot,
Gesture		iconic; face the parking lot	
Hand Combin		PT BHS	
RH Handshap		B_spread	
RH Orientatio		PTL FAB	
RH Position		CC D-CE LINE MF Medium	
LH Handshap		B_spread	
LH Orientatio		PTR FAB	
LH Position		CC D-CE LINE MF Medium	

Figure 8.4 The morphological features of the gesture shown in Fig. 8.3 are coded symbolically on an annotation window.

were taken of the dyad. For the videotape, four synchronized camera views were recorded (see Fig. 8.3). No time limits were imposed on the dyads.

8.3.2 Morphology Coding

The audio and video data was annotated in separate independent passes by a team of coders, using the *PRAAT*³ and *TASX Annotator* software (Milde & Gut, 2002). In the first pass, the words of the direction giver were transcribed. The next pass was for segmentation, in which the expressive, meaning-bearing phase of each gesture was spotted, and the gesture was classified according to the categories *iconic*, *deictic*, or *iconic+deictic* (all other types of gestures were ignored). In the final pass, the morphology of each included gesture was coded, using a scheme based on the McNeill Coding Manual (McNeill, 1992), refined for the purpose of our study. As shown in Fig. 8.4, the TASX annotator software was adapted to allow separate descriptions of the shape, orientation, and location of each hand involved in the gesture:

- *Hand shape* was denoted in terms of ASL (American Sign Language) shape symbols, optionally modified with terms like “loose”, “bent”, “open”, or “spread”.
- *Hand orientation* was coded in terms of the direction of an axis orthogonal to the palm, and the direction the fingers would point in if they were extended (see Fig. 8.5). Both were coded in terms of six speaker-centric, base- or half-axes (Herskovits, 1986a), namely forward, backward, left, right, up and down. Assuming the left hand in Fig. 8.5 is being held straight out in front of the body, it would be described as having extended finger direction forward (away from the body) and palm facing left. Combinations of these features were used to code diagonal or mixed directions (e.g. forward and to the left).
- *Hand location* was described relative to a zoning of the space in front of the gesturer as suggested by McNeill (1992), which determines a position in the frontal body plane. An additional symbol was used to denote the

³ <http://www.fon.hum.uva.nl/praat/>

distance between the hand and the body (in contact, between contact and elbow, between elbow and knee, or outstretched).

Movement in any of the three features was described using symbols to denote its shape (line, arc, circle, chop, or wiggle), its direction, and its extent. Again, directions were restricted to the six cardinal directions in space, or pairwise combinations of them. In addition to the three features for each hand, two-handed configurations (e.g., palms together, or finger tips of one hand touching the palm of the other) as well as movements of one hand relative to the other were explicitly denoted (e.g., hands move as mirror images, or one hand is held to anchor a frame of reference while the other hand is active). Figure 8.4 shows the results of morphology coding for the gesture in Fig. 8.3: The palms are touching each other (“PT”); the hands are mirror images, yet moving in the same direction (“BHS”); the hands are shaped flat (ASL shape “B_spread”); the fingers are pointed away from the body (“FAB”); the palms are facing toward right/left (“PTR”, “PTL”); the hands are positioned at the center of gesture space (“CC”), at a distance between contact and elbow (“D-CE”); a linear movement forward of medium extent is performed (“LINE MF Medium”). Gesture morphology was coded for the first four minutes or more of ten dyads, giving a total of 1171 gestures. The coded part of each dyad contains at least the descriptions for the first segment, i.e. between the same two buildings for all dyads (but with possibly different paths between them).

Morphology Coding Assessment

It is of course important in any investigation into the relationship between speech and gesture to be certain that the analysis is not circular, and that coding is rigorous. As far as the first question is concerned, as described above, speech and gesture were coded independently so that the content of the speech did not influence the coding of the morphology of the gesture. In addition, the coding of the morphology was carried out independently from the coding of the referents in the real world (see Section 8.3.3), such that neither influenced the other.

As far as the second question is concerned, inter-rater reliability is extremely hard to assess for coding in which there are as many as 12 sub-parts to each coding decision (extent of gesture, kind of gesture, shape of right hand, trajectory of right hand, shape of left hand, location in space of left hand . . .). We therefore depended on two methods to ensure rigor, and to assess accuracy. First of all, all coding, both of gesture and of referents in the world, was carried out by a minimum of two coders, with any disagreements resolved by discussion. And accuracy was assessed by asking four subjects who played no role in coding to *reproduce* 75 randomly chosen gestures from the dataset, solely on the basis of the codes (one subject 15 gestures, three subjects 20 gestures each). All of the subjects were members of our lab, i.e. generally familiar with gesture, but none had seen any of the original movies or dealt with this dataset. After being trained to our coding manual in a short practice session (five gestures, solely verbal instructions), each subject was videotaped while reproducing the test gestures (coding manual available to them to interpret abbreviations). These video recordings were then compared with the original data to assess similarity between the original gestures and the reproduced ones. Similarity was rated from 1 (identical) to 4 (completely different), separately for hand shape, hand orientation, and hand location (plus movement), based on criteria specific to each feature (Table 1 lists the criteria).

Hand shape	1: same shapes 2: same shape, but wrongly modified (e.g. 'open'); 3: different, but similar shape (e.g. ASL ‘5’ and ‘B’); 4: otherwise
Hand orientation	1: same orientation; 2: extended finger directions or palm directions differ less than 45°; 3: directions differ more than 45°, but less than 90°; 4: otherwise
Hand position	1: same position; 2: different, but still in the same gesture space region (according to McNeill) and distance; 3: in adjacent regions; 4: farther away
Movement	1: same movement; 2: same direction and plane of movement, but slightly different extent or shape (e.g. arc, but stronger curved); 3: movements differ considerably in either shape, extent, direction, or plane of movement; 4: movements differ considerably in at least two of the four criteria

Table 8.1 Criteria used for reproducibility of morphological coding.

Three gesture were excluded for analysis, since they were performed with the wrong hand, leaving a total of 72 gestures, of which 15 were static and 57 included movement (19 linear, 23 arcs, 10 chop-like, 3 circles). Similarity in each feature was judged for each of these gestures independently, and then the arithmetic mean was calculated. The resulting average value across all gesture ratings was 1.54 (SD=0.44), with static gestures being reproduced

more accurately than gestures which include movement. Table 2 shows the results for the different features, for each kind of gesture. Overall, the results of the recreation assessment test indicate that the morphology codes specify almost all of the information needed to reproduce a gesture that is almost identical to the original. Even more importantly, the number of errors in our form coding proves to be well within acceptable limits.

	Aver.	Static	Dynamic	LINE	CHOP	ARC	CIRCLE
Hand shape	1.28	1.54	1.20	1.15	1.06	1.36	1
Orientation	1.42	1.34	1.45	1.15	1.26	1.53	2
Position	1.86	1.34	2.04	1.69	1.93	2.4	2
Average (SD)	1.54 (0.44)	1.43 (0.26)	1.57 (0.47)	1.33 (0.38)	1.52 (0.46)	1.73 (0.51)	1.54 (0.44)

Table 8.2 Quality of morphology coding, values range from 1 (correct) to 4 (completely different).

8.3.3 Referent Coding

Using independent information about the campus from maps, photographs and trips across campus, three of the experimenters named each place on campus that was referred to by the first four minutes of each of the complete set of direction-giving episodes. Those episodes yielded 195 unique, concrete objects or subparts of objects, including parking lots, signs, buildings, lakes and ponds, etc.

Finally, we chose several visuo-spatial features to investigate based upon Landau & Jackendoff (1993) and Talmy (1983), and marked each referent as to whether or not it we perceived the feature as salient when looking at it from various angles in the route perspective (i.e. not from above). Paths and roads were marked as "sideways planes", which refers to objects which have to parallel sides or borders that seem to be conceptualized as one dimensional lines or two dimensional planes. This last category corresponds to the "ribbonal" plane, or "a plane bounded by two parallel edges" defined by Talmy (1983). In general, every building was marked as having vertically oriented plane or surface features, corresponding to its walls. Every lake and parking lot was marked with horizontal planes, corresponding to the surface of region. Many subparts of buildings were referred to, e.g. windows, which were marked with vertical planes, likewise for various kinds of signs. A few buildings were also marked with horizontal plane features, for example Northwestern University's Central Utility Plant, which is a low, long building; when one walks by the building, the surface of the roof is a salient characteristic.

8.3.4 Pre-study

In order to determine the feasibility of the approach, in a preliminary analysis (Kopp, Tepper, & Cassell, 2004) we selected several hundred gestures illustrating several particular morphological features, and combinations thereof, and evaluated what they referred to. The results suggested a correspondence between combinations of morphological features and the visual or geometrical similarity of referents. For example, we found that 67% of the gestures with a linear trajectory (N=48) referred to objects, and that the gestures tended to depict a significant axis with the linear movement (e.g., length of a street, extent of a field, transverse overpass). Likewise, 80% of gestures with a flat hand shape and the palm oriented vertically (N=45) referred to objects whose shape comprised an upright plane (walls, stop sign, window, etc.).

85% of the gestures with a flat hand shape and the palm oriented sideways (N=61) referred to directed actions (go, keep on, take a, look, etc.), and they always finished with the fingers pointing in the direction of the action. These gestures seemed to fuse iconic with deictic aspects, in that the trajectory of the hands depicted the concrete path or direction of the action described, while hand shape and palm orientation appeared to be more “conventionalized” in their indexing of locations and directions. These results gave us the confidence to continue the analysis on the full set of 1000 gestures.

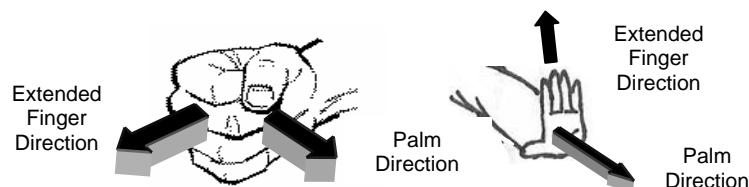


Figure 8.5 Hand Orientation defined in terms of Extended Finger Direction and Palm Direction.

8.3.5 Analysis of Morphology Features and Visual Features of Gesture Referents

Our first analysis looks at simple morphological configurations to see if they correlate with two particular IDF-sets. The morphological configurations comprise combinations of hand shape features that appear relatively “flat” (e.g., “5”, “B”, and their loose and open variants, according to the ASL alphabet), in several orientations in space. That is, we looked at flat hand shapes oriented vertically (with fingers pointing up), and horizontally (with palm pointing down) to see if they correlated positively with referents that possess salient visual characteristics corresponding to vertical and horizontal planes. Again using maps, photographs and trips across campus, we annotated each landmark (including landmark aspects and subparts) with information on whether it had these salient characteristics. For example, the walls of a tall building contain vertically oriented planes (surfaces) and a parking lot contains a horizontal plane (surface). These two features could be formalized using IDFs to relate them to gesture morphology. For example a vertical plane might be formalized as: *building(ams) ∧ has_part(wall1,ams) ∧ isa(wall1,surface) ∧ orientation(wall1, vertical)*. It is important to note that these features are not mutually exclusive. For example, some buildings seem to possess both vertical surfaces (walls) and horizontal planes (their wide footprint, overhanging roof, etc.). Figure 8.6 illustrates (from left to right) the kind of correspondence we hypothesized, between a gesture with a flat, vertically oriented morphology (cf. Fig. 8.5), a vertical plane (surface) definable in terms of IDFs, and a building (wall) that has the vertical plane feature.

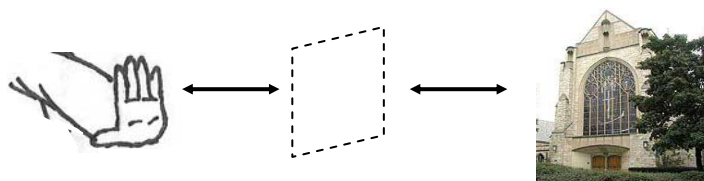



Figure 8.6 (From left to right) Gesture morphology, corresponding salient visual characteristic (IDFs) and a corresponding concrete referent.

Because gestures do not fall neatly or naturally into such discrete categories, various levels of strictness for their high-level categorization were chosen. Specifically, for deciding whether a gesture should be categorized morphologically as having a flat shape plus a vertical or horizontal orientation, we defined a *weak* and a *strong* version for each of our sets of criteria. Since we have two classes of orientation features—finger and palm—we required a 2×2 classification system for each feature, as shown in Table 8.3.




Palm Dir.	Extended Finger Direction	
	Weak	Strong
Weak	F = Up or Down	F = Up or Down F ≠ Left, Right, Forward or Backward
Strong	F = Up or Down P ≠ Up or Down	F = Up or Down F ≠ Left, Right, Forward or Backward P ≠ Up or Down

Table 8.3 Flat hand + Vertical orientation + Extended-finger-direction up or down


Each cell of this table contains a set of criteria, or constraints. These specify a set of variations on a gesture with a flat hand shape, extended finger direction up or down, and palms facing any direction but up or down. Abbreviating extended finger direction as *F*, palm direction as *P*, weak as *w* and strong as *s*, the box for *FwPw* contains the constraint “*F* = Up or Down”, meaning that the extended finger direction must have an up or down aspect. This can be combined with other features, e.g. it can be up and to the left, overall a weak version of pointing up. The strong *F* cells include additional constraints, prohibiting the inclusion of the left, right, forward or backward features; i.e. the *F* feature is purely up or down.

In looking at these features, it turned out that a simple “flat vertical” classification was not specific enough. Vertically-oriented hand shapes seemed to be used in meaningfully different classes of gestures that differ in extended finger direction, where *vertical* means that the extended finger direction is up or down; *horizontal* means that the fingers are pointing left, right, forward or backward, with the palm oriented downwards; and *sideways*, where the palm is oriented towards the left or right, and the thumb is pointing up or down. Table 8.3 shows the criteria for flat vertical, where extended finger direction is always up or down, Table 8.4 for flat sideways. Similarly, Table 8.5 shows the criteria for flat “horizontal” morphology, resembling a horizontal plane.



Palm Dir.	Extended Finger Direction	
	Weak	Strong
Weak	F = (Left or Right and/or Forward or Backward) P = (Left or Right and/or Forward or Backward)	F ≠ Up or Down P = (Left or Right and/or Forward or Backward)
Strong	F = (Left or Right and/or Forward or Backward) P ≠ Up or Down P = (Left or Right and/or Forward or Backward)	F ≠ Up or Down P ≠ Up or Down P = (Left or Right and/or Forward or Backward)

Table 8.4 Flat hand + Vertical orientation + Extended Finger direction forward, backward, right or left



Palm Dir.	Extended Finger Direction	
	Weak	Strong
Weak	P = Up or Down	F ≠ Up or Down P = Up or Down
Strong	P = Up or Down P ≠ Left or Right P ≠ Forward or Backward	F ≠ Up or Down P = Up or Down P ≠ Left or Right P ≠ Forward or Backward

Table 8.5 Flat hand + Horizontal orientation

With these specifications in place, we can now turn to the questions we tried to answer with this analysis: What is the degree of correspondence between the occurrence of a shape feature of the referent of a gesture and the occurrence of a morphological feature in the gesture? More specifically, which of the four levels of classification stringency provides the highest degree of correspondence between the shape features of the referents and the shape features of the gestures? We expect that the less stringent levels will be the weaker predictors, and the more stringent levels will be stronger predictors. Although we suggest that the features of the landscape determine the features of the gesture and not the other way around, there are potentially infinite features of referents about which to gesture; therefore, it is impractical to examine whether features of the referents predict features of the gestures. Accordingly, we examine instead the extent to which shape features of gesture morphology predict the existence of corresponding features in the referent. We use dichotomous outcome decision tables (an effective tool used in evaluation research), with gestures classified as possessing horizontal plane, vertical plane, and sideways plane features or not at our four levels of stringency (12 decision tables in total). Table 8.6 shows the decision table for the *FwPw* (Finger-weak Palm-weak) classification of the gestures with flat hand shapes, vertical orientation, and extended finger direction up or down. (We will abbreviate this set of morphological features *flat-vertical*.)

FwPw	Occurrence of a Vertical Plane in the Referent	No Occurrence of a Vertical Plane in the Referent	
Flat Vertical Gesture Morphology (Predicted occurrence of vertical plane in the referent)	77	46	123
No Flat Vertical Gesture Morphology (No predicted occurrence of vertical plane in the referent)	137	178	315
	214	224	438
Table 8.6 Example of a frequency table used in this study.			

In this case, 77 gestures of the 123 that have the flat vertical morphology also refer to a landmark that has the predicted vertical plane feature, while the other 46 do not. Of the 315 gestures that do not have flat vertical morphology, 178 also do not refer to a landmark with the vertical plane feature. The referents of the other 137 gestures have a vertical plane feature. We calculated the chi-square value and the phi coefficient of the data in this and the rest of the decision tables used in this study. The chi-square *p*-value indicates the likelihood that our results occur by chance, and the phi coefficient is a measure of the correlation between our predictors and outcomes, or the validity of our gesture morphology as a predictor.

8.3.6 Results

Our results are summarized in tables 8.7 – 8.9. Of the three gesture features we examine, the horizontal morphology is the weakest predictor overall, while the Finger-strong Palm-weak stringency level provides the best predictions across the three gesture morphology types. Upon further examination of the horizontal morphology predictions, we find that about one fifth of the horizontal morphology gestures depict buildings, which do not have the horizontal plane feature. We'll return to this point in the discussion. The level of stringency of flat-horizontal gestures that provides the best prediction of horizontal planes in the referent is the Finger-strong Palm-weak level, with a small but statistically significant validity (Φ) of .1194. That is, when the fingers are not pointing up or down at all, but the palm is not necessarily perfectly horizontal, we are most likely to find that the referent contains a salient horizontal plane.

Table 8.7 Percentage of Flat **Horizontal** Gestures depicting Referents with Horizontal Planes and other features (numbers in parentheses)

Stringency of Morphology Categorization	Referent Feature			Total	χ^2	Test Validity (Φ)
	Horizontal Plane	No Horizontal Plane				
		Building	Misc. Referent			
FWPW	50.6% (49)	21.6% (21)	27.8% (27)	100% (97)	3.1153	.0843
FSPW	57.4% (35)	18.0% (11)	24.6% (15)	100% (61)	6.2449*	.1194*
FWPS	55.6% (25)	20.0% (9)	24.4% (11)	100% (45)	3.3909	.088
FSPS	55.6%(25)	20.0% (9)	24.4% (11)	100% (45)	3.3909	.088

* $p < .05$

For the prediction of the existence of vertical planes in the referent by occurrence of flat-vertical morphology in the gestures, we find that all levels of stringency are statistically significant. The flat-vertical morphology is the strongest predictor overall. Within flat-vertical morphology, the Finger-strong Palm-weak stringency level is the best predictor, with a somewhat larger validity (Φ) of 0.2071, significant at the .001 level. Note also that the Finger-strong Palm-weak and Finger-strong Palm-strong stringency levels return the same number of gestures.

Table 8.8 Percentage of Flat **Vertical** Gestures depicting Referents with Vertical Planes and other features (numbers in parentheses)

Stringency of Morphology Categorization	Referent Feature		Total	χ^2	Test Validity (Φ)
	Vertical Plane	No Vertical Plane			
FWPW	63% (77)	37% (46)	100% (123)	12.9279*	.1718*
FSPW	79% (37)	21% (10)	100% (47)	18.7934*	.2071*
FWPS	68% (59)	32% (28)	100% (87)	15.615*	.1888*
FSPS	79% (37)	21% (10)	100% (47)	18.7934*	.2071*

* $p < .001$

Finally, we find that, for prediction of the existence of ribbonal planes in the referent by occurrence of flat-sideways morphology in the gestures, all levels but the weakest are statistically significant. The Finger-strong Palm-weak stringency level is again the most effective of the four levels, with a validity (Φ) of .2002, significant at the .001 level.

Table 8.9 Percentage of Flat_Sideways Gestures depicting Referents with Sideways Planes and other features (numbers in parentheses)

Stringency of Morphology Categorization	Referent Feature		Total	χ^2	Test Validity (Φ)
	Sideways Plane	No Sideways Plane			
FWPW	46% (100)	54% (117)	100% (217)	2.0188	.0679
FSPW	56% (86)	44% (67)	100% (153)	17.5546***	.2002***
FWPS	58% (28)	42% (20)	100% (48)	5.3891*	.1109*
FSPS	61% (28)	39% (18)	100% (46)	6.9399**	.1259**

* $p < .05$; ** $p < .01$; *** $p < .001$

8.4 Discussion of Empirical Results

Did we find evidence for a systematic relationship between the form of gestures and their meaning -- between the visual characteristics of the gestures and the spatial features of the entities they refer to? Yes we did, for every morphological class that we investigated. Of the three morphological classes we explored, the vertical flat morphological form was the best predictor of referent form. However, for all three morphological classes, a significant correlation exists between a “more-or-less-flat” handshape with perfectly-oriented extended finger direction (*FsPw*) and the spatial features of the referent.

There are, however, several caveats. First of all, it should be noted that there are quite a number of false negatives in each frequency table. That is, although it is frequently the case that a flat vertical handshape refers to a flat vertical image description feature linked to, for example, a tall building, it is also quite often the case that a tall flat building is not described using a flat handshape. From the perspective of gesture analysis, this is quite understandable. Route directions include reference to landmarks as a way of ensuring that the listener is on the correct path. It therefore stands to reason that landmarks must be described in such a way as to disambiguate them from the other buildings and objects that are nearby. If this is the case, then gesture may participate in the act of disambiguating a referent from a kind of distracter set – a quite different task than simply accomplishing reference to a building in and of itself. In order to determine whether gesture is serving this function, we would need to go back and code the referents for the visuo-spatial features that serve to differentiate them from their neighbors, rather than the features that are generally visually salient.

Secondly, the results did not – as we had expected – adduce evidence for stronger morphological features being more strongly linked to features of referents. The effect of strictness criteria for palm and extended finger direction is significant, but a clear pattern of predictive power does not emerge. In fact, for every morphological class, the most predictive stringency level was strong fingers and weak palm. This result may derive from the different ways in which gestures are produced depending on (a) the response of the listener, and (b) the discourse function of the entity being referred to, the placement of the reference in the set of directions. That is, we might expect gesture to be made more forcefully the first time an entity is introduced, or when an entity is more important, or when the listener shows signs of not following. In a sense, weak features may act like proforms – like the schwa in “the” – once a referent has been established with a stronger handshape. This hypothesis certainly bears further investigation.

Finally, we coded all of the referents as individual entities. Future work should attempt to identify systematicity in the features regarded as salient for each referent or referent class. Thus, buildings may always be portrayed in gesture with horizontal flat features – a sort of lexicalization of gesture. This could be carried out by examining the correlations between morphological configurations within a group of gestures that all refer to the same thing, e.g., a particular landmark. At this point, we do not have enough data to explore this question.

In our virtual human system, this information about referent classes is needed for use in content planning of natural language and gesture, which selects the IDFs and semantics for inclusion in the goals it sends to the

microplanner. But for microplanning we only needed evidence of the existence of correlations—systematicity—which we have found. And so, with this evidence in hand, we now turn to the task of modeling direction-giving with embodied conversational agents.

8.5 Generating Directions with Humanoids

Based on our empirical results, we can now approach one of the problems in generating directions for an ECA: the encoding of spatial and visual information into appropriate words and iconic gestures. Although much existing work has addressed the automatic generation of coordinated language and visualization for complex spatial information (Towns, Callaway, & Lester, 1998) (Kerpedjiev, Carenini, Green, Moore, & Roth, 1998) (Green, Carenini, Kerpedjiev, & Roth, 1998), little of this research has addressed coordinated generation of speech and gesture for spatial tasks.

Traum & Rickel (2002) present a model of dialogue acts for spoken conversation that incorporates non-verbal behavior into its representation as well as accounts for a representation of the discourse state of these dialogue acts. This work is related in that it deals with discourse state of non-verbal behavior (Rickel et al., 2002), but it does not consider questions of generating these behaviors. Nijholt et al. (Nijholt, Theune, & Heylen, 2005) discuss architectural issues for multimodal microplanning and the factors influencing modality choice, but adhere in their proposed model to selecting iconic and deictic gestures from a lexicon; the issues of iconicity of gesture and their underlying semantics are not considered. To date, the *REA* system (Cassell, Stone, & Yan, 2000) represents the most elaborated work on the automatic generation of natural language and gesture in embodied conversational agents (ECAs). Using the SPUD system (Stone, Doran, Webber, Bleam, & Palmer, 2003) for planning natural language utterances, *REA* was able to successfully generate context-appropriate language and gesture, relying upon empirical evidence (Cassell & Prevost, 1996; Yan, 2000) that communicative content can be defined in terms of semantic components, and that different combinations of verbal and gestural elements represent different distributions of these components across the modalities. This approach was able to account for the fact that iconic gestures are not independent of speech but vary with the linguistic expression they accompany and the context in which they are produced, being sometimes redundant and sometimes complementary to the information conveyed in words. However, whole gestures were treated exactly like words, associated to syntactic trees by a specific grammatical construction, the SYNC structure, and gesture planning only extended as far as the selection of a complete gesture from a library and its context-dependent coordination with speech. This does not allow for the expression of new content in gestures, as is possible in language with a generative grammar. Gao (2002) extended the *REA* system to derive iconic gestures directly from a 3D graphics scene. He augmented the VRML scene description with information about 3D locations of objects and their basic shapes (boxes, cylinders, spheres, user-defined polygons, or composites of these), which were mapped onto a set of hand shapes and spatial hand configurations. This method allows for deriving a range of new gesture forms, but it does not provide a unified way of representing and processing the knowledge underlying coordinated language and gesture use.

The fact that previous systems usually draw upon a “gestionary”, a lexicon of self-contained gestures, is also a consequence of the use of canned gesture animations. Although previous systems, e.g. *BEAT* (Cassell, Vilhjálmsón, & Bickmore, 2001), were able to create nonverbal as well as paraverbal behaviors—eyebrow raises, eye gaze, head nods, gestures, and intonation contours—and to schedule those behaviors with respect to synthesized text output, the level of animation was always restricted to predefined animations. Sometimes, motor primitives were used that allowed for some open parameters (e.g., in the *STEVE* system (Rickel et al., 2002) or *REA* (Cassell, Stone, & Yan, 2000)), were adjustable by means of procedural animation (*EMOTE* (Chi, Costa, Zhao, & Badler, 2000)), or could be combined to form more complex movements (e.g. Perlin & Goldberg, 1996). For example, Kopp and Wachsmuth (2004) presented a generation model that assembles gestural motor behaviors on the fly, entirely based on specifications of their desired overt form. This method allows for greater flexibility with respect to the producible forms of gesture, which is clearly a prerequisite for the level of gesture generation targeted here, but it does not determine the morphology of the gesture from the communicative intent.

In the current work, by following the patterns between IDFs and form features, we can plan a detailed morphology of a gesture from a given communicative intention. This process takes place in the context of simultaneous speech, and the gesture will be underspecified until it is interpreted in concert with the accompanying words. In our approach, we extend a Natural Language Generation (NLG) model to the integrated generation of both natural language and iconic gesture (henceforth, NLGG). Commonly, NLG systems have a modular, pipeline architecture, broken down into three subtasks—content planning, microplanning and surface realization (in that order, Reiter & Dale, 2000). In ordinary language, the work done by these three subsystems boils down to,

respectively, figuring out what to say, figuring out how to say it, and finally, saying it. In this article we focus on microplanning, the second stage of the NLGG pipeline, where domain knowledge must be recoded into linguistic and gesture form, although we must first outline some prerequisites to be met by the other stages for this.

8.5.1 Modeling Content and Spatial Information

The content planning selects and structures domain knowledge into coherent directions. We do not discuss this process here as it is beyond the scope of this paper (but see e.g. Guhe, Habel, & Tschander, 2003; Young & Moore, 1994). However, we must note that our NLGG model requires a rich representation of domain knowledge that pays attention to the affordances of both language and gesture as output media. In our present project on direction giving, most of this content is spatial information about actions, locations, orientations, and shapes of landmarks, and we need a representation powerful enough to accommodate all information expressible in spatial language and gesture. To model spatial language, we require two levels of abstraction, with corresponding layers of formal representations. For example, for a NLG system to refer to an object as “tall”, first, the concept or property of tallness must be formalized. This can be done as a simple logical formula like $tall(X)$, where $tall$ is a predicate symbol representing the concept, and X is an open variable which can be bound to another ground symbol, representing a particular discourse referent (e.g., $tall(church)$ or $tall(john)$). Second, this formula must be associated with the string “tall” representing the word itself. This level of granularity is too coarse for iconic gesture, for which a more fine-grained specification in terms of the intrinsic spatial nature of this property is required. For example, tallness can be described as holding of an object when the extent of its vertical axis is longer than its other axes, or more likely it is long relative the vertical axes of some other relevant objects (e.g., a man might be tall relative to some other men standing nearby), or relative to some stereotype. We use the intermediate IDF level to represent such spatial properties that can be displayed by gesture. If the concept of tallness is represented as $tall(X)$, and its spatial description is represented as a set of IDFs, we can then map these IDFs onto form features, and this iconic gesture can be used to refer to the concept.

This example motivates IDFs and conceptual/semantic knowledge as different kinds of knowledge with different levels of abstraction, needed to exploit the representational capabilities of the two modalities, and meriting separation into two ontologically distinct levels. However, at the same time, we follow the ideas of one amodal, common representation of—even spatial—content (e.g. Landau & Jackendoff, 1993) that both language and gesture utilize, working together to express information as parts of one communicative system (McNeill, 1992). We thus maintain a single, common representation system encompassing all the kinds of domain knowledge needed, formalized in terms of qualitative, logical formulae. We base this system on a formal, extensible ontology that encompasses objects (buildings, signs, etc.), regions (parking lots, lake, etc.), and actions (go, turn, etc.). In addition, it defines IDF-related symbols for basic shapes, locations, directions, or qualitative extents (long, short, large, tall, narrow, etc.). When ontologically sound, entities are connected using taxonomic (is-a), partonomic (part-of), and spatial relations (in, on, left-of, etc.). The ontology thus provides for IDFs and lays down their assignment to concrete objects or actions. In other words, it provides for the link between an entity and a mental image thereof—the latter formalized in terms of IDFs. We have built such an ontology for parts of Northwestern University campus, including all landmarks that were referred to in the analyzed route descriptions. Content plans are then specified in terms of these entities and relations. Figure 8.7 shows an example content plan that comprises all kinds of knowledge—including IDFs—required for employing language and gesture in instructing someone that she will see a particularly shaped building (“Cook Hall”) on her right.

```
instruction(e2). see(e2,user,cook,future,place(on,right)). tense(e2,future).
name(cook,cook_hall). type(cook,building). place(on,right).
rel_loc(cook,user,right). shape(dim,vert,cook).
shape(primary_dim(longit,cook)). shape(dim,longit,cook).
```

Figure 8.7 Content plan in logics notation (propositions are delimited by points).

8.6 Multimodal Microplanning

The multimodal microplanner must link domain-specific representations of meaning, just like the ones shown in Fig. 8.7, to linguistic form and gesture form. As language and gesture require different kinds of information, provide different representational capacities, and convey information in different ways, NLGG calls for specific models of

how each modality encodes content. We thus add a new subsystem to the microplanning stage of NLGG, as illustrated in Figure 8.8. This new component, the gesture planner (GP), is responsible for planning the morphology of a gesture appropriate to encode a set of one or more input IDFs. That is, the GP is itself a microplanner, addressing the problem of recoding content into form, but this time on a feature level, from IDFs to morphological features.

To connect content to linguistic forms, we employ a grammar-based sentence planner, SPUD (Stone, Doran, Webber, Blear, & Palmer, 2003). SPUD takes a uniform approach to microplanning, framing it as a search task wherein utterances are iteratively constructed from an input specification of a set of resources and a knowledge base (KB) that contains, among others, the facts to be communicated (communicative effects). All facts are explicitly labeled with information about their conversational status, e.g. whether the fact is private or shared, constraining decisions about what information the system must assert as new to the user, and what it can presuppose as information in the common ground (Clark, 1996). The grammar includes a set of lexical entries and a set of

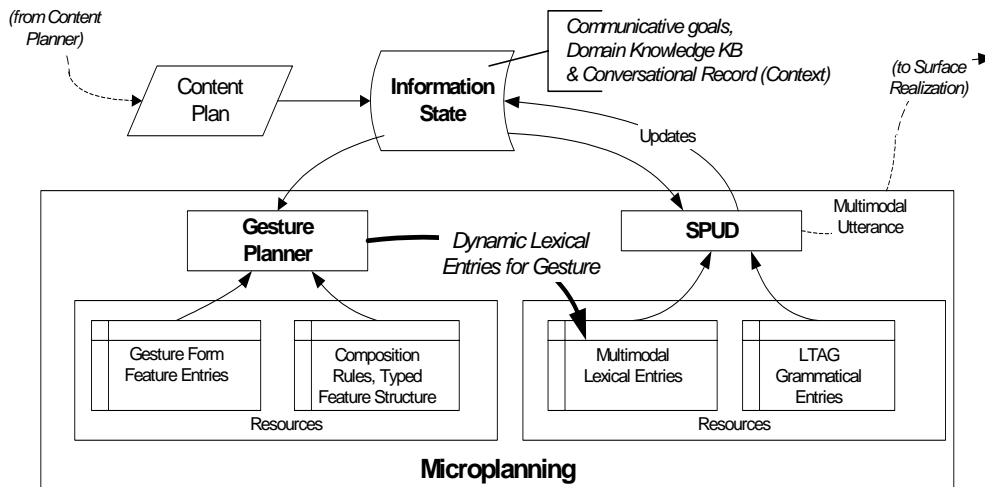


Figure 8.8 Overview of the multimodal microplanning process.

syntactic constructions, formalized using Lexicalized Tree Adjoining Grammar (Joshi, 1987), and optionally containing pragmatic constraints on use of this construction relative to discourse context. Each lexical entry consists of a lexical item (word), a set of logical formulae defining its semantics and, again, pragmatic conditions for use in conversation. SPUD works towards a complete, grammatical structure by iteratively selecting words and syntactic structures to add, such that their semantics allows for the assertion of the maximum number of communicative effects to be achieved per state (simulating an economical, Gricean approach to generation). Additionally, for each state, the system maintains a representation of the utterance's intended interpretation, or communicative intent, a record of inferential links made in connecting the semantics and pragmatics, associated with linguistic terms, to facts about referents in the world, as recorded in the KB. In our current project, we use a fast, lightweight, Prolog implementation of SPUD, wherein inference from open variable parameters to particular referents in the KB is achieved via Prolog unification.

In previous work (Cassell, Stone, & Yan, 2000), SPUD's linguistic resources have already been extended to include a set of predefined gestures, from which it drew upon to express its communicative goals. We follow this same strategy here, using SPUD to compose full, multimodal utterances via a single, uniform algorithm. But, instead of drawing upon a static set of predefined gestures, we add the GP into the pipeline: before calling SPUD, the GP plans iconic gestures that express some or all of the given IDFs. The planned gestures are then dynamically incorporated into SPUD's (now multimodal) resources and utilized in the same way as described in (Cassell, Stone, & Yan, 2000).

8.6.1 Gesture Planning and Integration

Similar to the sentence planner SPUD, the Gesture Planner system draws upon a bipartite input specification of domain knowledge, plus a set of entries to encode the connection between semantic content and form. Using such data structures, we are able to achieve the same kind of close coupling between gesture form and meaning, allowing for efficient, incremental construction of gestures and maintenance of inferential links from abstract meaning

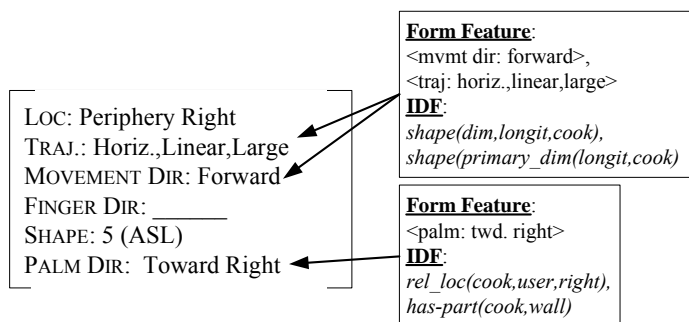


Figure 8.9 Example of form features entries filling a gesture feature structure.

(logical) formulae to specific discourse referents. For the GP, we formalize the form-meaning coupling in a set of “form feature entries”, data structures that connect IDFs to morphological features. Form feature entries implement the patterns that we found in our empirical data, and may contain “clusters” of features on either side, i.e., conjunctions of IDFs as well as combinations of morphological features. Also, we use these entries to encode the ways in which the function of a gesture (e.g., deictic) influences its morphology (e.g., hand shape) through conventionalized patterns, again, as suggested by our empirical data.

When receiving a set of IDFs as input, the desired communicative effects, the GP searches for all combinations of form feature entries that can realize them. Because, as we have seen, iconic gesture is not governed by a hierarchical system of well-formedness, we employ feature structure unification to combine morphological features, whereby any two features may combine provided that the derived feature structure contains only one of any feature type at a time. Through iterative application of this operation, the GP builds up gestures incrementally until all the desired communicative effects are encoded. Figure 8.9 shows a state in the generation of a gesture, composed to depict the IDFs from the content plan in Figure 8.7. Location and hand shape have already been inserted, the latter according to one pattern we observed in our data, namely, the use of a flat hand shape (ASL sign 5) and a vertically oriented palm for depicting the wall of the Cook building. This pattern now informs the palm orientation, together with the location of the object (cook) to be depicted. Note that the GP may output an underspecified gesture if a morphological form feature does not meaningfully correspond to any of the disposed IDFs, i.e., it remains undefined by the selected patterns

Similar to SPUD’s pragmatic constraints on the way language is used in context, the GP process can be guided by composition constraints on all possible ways to combine a set of form features into a feature structure that defines a realizable gesture. Such composition constraints could formalize restrictions over the ways in which different form features combine, and could, for example, be utilized to favor the reuse of feature structures that have been successfully employed before to express a common set of semantic formulae. This would require comparison to the KB’s record of context, and allows for simulation of what McNeill (1992) has called catchments, the maintenance of a certain gesture morphology to indicate cohesion with what went before.

In our current Prolog implementation, the GP returns all possible combinations of morphology features, i.e., it delivers all gestures that could take on communicative work by encoding some or all of the desired communicative effects. Each dynamically planned gesture is added to SPUD’s resources, which also contain a set of dedicated SYNC constructions which state the possible ways in which language and gesture can combine. Each SYNC construction pairs a certain syntactic constituent and a gesture feature structure under the condition that their predicate arguments are connected to the same discourse referents, achieving coordination of meaning in context. In addition, it imposes a constraint of temporal surface synchrony between both elements. The SPUD algorithm chooses the gesture feature structure and the construction that, when combined with appropriate words, allow for the most complete intended interpretation in context (see (Cassell, Stone, & Yan, 2000) for details). Figure 8.10 shows how a gesture feature structure, derived for the IDFs in the content plan form Figure 8.7, is combined with a linguistic tree to form a multimodal utterance. Finally, the tree of the resulting multimodal utterance is converted into an XML description, containing the textually defined words along with the feature structures for gesture. This tree is passed on to the next and final stage of our NLGG pipeline, surface realization.

8.7 Surface Realization

Starting out with the XML specification outputted by the Microplanner, Surface Realization concerns turning this tree into multimodal output behaviors to be realized by our embodied conversational agent NUMACK (the *Northwestern University Multimodal Autonomous Conversational Kiosk*). This is done in two steps, Behavior Augmentation and Behavior Realization.

8.7.1 Behavior Augmentation

The XML specification coming in from microplanning only amounts to communicatively intended behaviors, i.e., words and expressive gestures directly derived from communicative goals. To achieve a more natural, multimodal output, the Surface Realization engine has to impart to the utterance additional nonverbal and paraverbal behaviors like intonation, eyebrow raise, head nod, or posture shifts. These behaviors do not encode explicit communicative goals, and are thus not conceived during microplanning, but are essential to meaning as they underline the conveyance of central parts of the utterance.

For this task, we employ the *BEAT* system (Cassell, Vilhjálmsón, & Bickmore, 2001). Its approach is to suggest all plausible behaviors first, and then use filters to trim these over-generated behaviors down to a set appropriate for a particular character, all carried out on the same XML tree. Behavior suggestion draws upon information about grammatical units (clause structure), information structure (theme and rheme), word newness, and contrast, each of which represented by dedicated tags in the XML tree. This tree gets augmented with appropriate behaviors by applying each of an extensible set of rule-based generators to all XML nodes. When a node meets criteria specified by a generator, a suggestion is added independent of any other by inserting a behavior node. Its position in the tree defines the time interval the behavior is supposed to be active; namely, synced with all words contained in its subtree. Currently, we employ such generators for gaze, intonation, eyebrow raises, and head nods; see (Cassell, Vilhjálmsón, & Bickmore, 2001) for a more detailed description. Note that the information at disposal for behavior suggestion is limited to clause structure and information structure, the latter being always set to ‘rheme’, as our system currently lacks a discourse planner for composing multi-clause utterances with thematic parts.

Behavior selection, then, applies filters to the tree which delete all behavior suggestions that cannot physically co-occur or whose assigned priority falls below a pre-specified threshold. Figure 8.11 shows a Behavior Augmentation example, in which gaze and intonation behaviors are added to the utterance “*Make a left*”. In addition, the clause is turned into a “chunk” node which demarcates a unit for speech-gesture realization to follow.

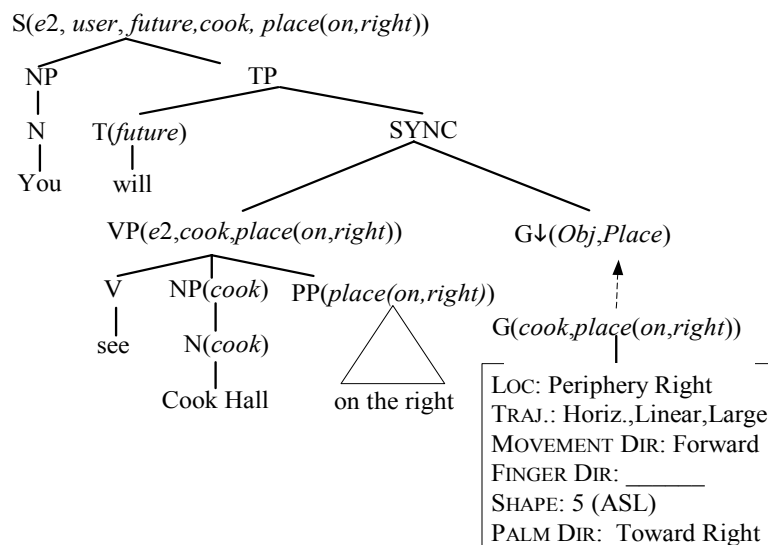


Figure 8.10 Insertion of the gesture into the utterance tree.

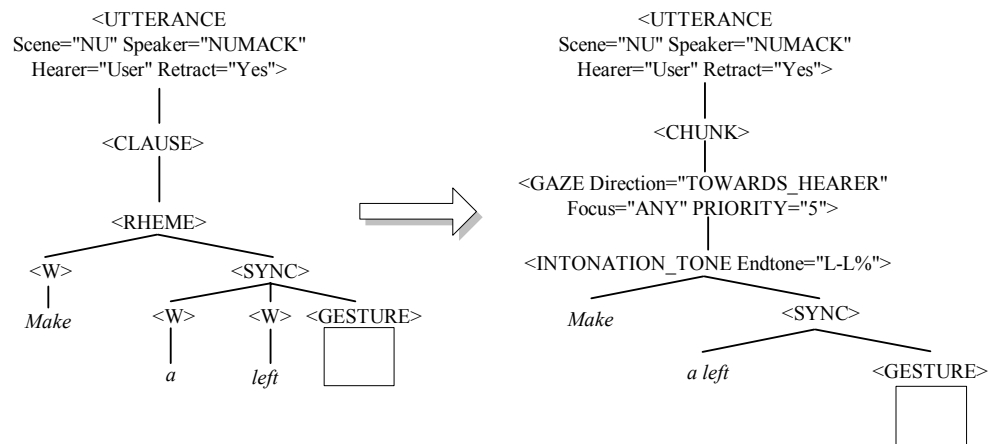


Figure 8.11 Insertion of non-/paraverbal behaviors in a multimodal utterance tree during Behavior Augmentation.

8.7.2 Behavior Realization

Upon completion of Behavior Augmentation, the XML tree with its morphologically and phonologically-specified behaviors is turned into synthesized speech and intonation, expressive gestures and other animations for a graphical avatar body, all being scheduled into synchronized multimodal output. We employ the *MAX* system (Kopp & Wachsmuth, 2004), which provides modules for synthesizing gestural, verbal, and facial behaviors, and embeds them in an incremental process model to schedule and link their deliveries into synchronized, fluent utterances.

Before Behavior Realization can start, the gesture description must first be converted into a format that can be processed by *MAX*'s gesture generation module. In the XML tree delivered by the microplanner, a gesture is stated as a typed feature structure that contains a possibly incomplete set of form features. As illustrated in Figure 8.12, this feature structure is now translated into *MURML* (Multimodal Utterance Representation Markup Language; Kopp, Tepper, & Cassell, 2004), a XML-conforming, feature-based representation that denotes a hand-arm configuration in terms of the same features as the original feature structure, but using a slightly different set of descriptive symbols based on *HamNoSys*, a notation system for German sign language (Prillwitz, 1989). A gesture is described as a combination of separate, yet coordinated postures or sub-movements within the features, the latter being defined as a sequence of elements each of which partially guiding it. To explicate the gesture's inner structure, features are arranged in a constraint tree by combining them with dedicated nodes for expressing simultaneity, posteriority, symmetry, and repetition of sub-movements (for example, the *PARALLEL* node in the tree in Figure 8.12 denotes simultaneity of its child features). Form features that have been left open by the microplanner are either set to default values, or remain undefined when no action needs to be taken about them. For example, a gesture whose location has not been laid down during microplanning is per default performed in the center of

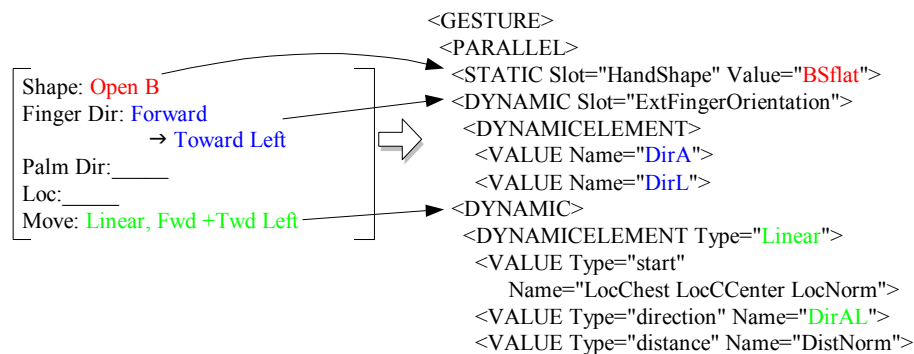


Figure 8.12: Conversion of a feature structure into a MURML specification.

gesture space.

The complete multimodal utterance tree is finally processed by the realization model that carries out and coordinates scheduling, synthesis, and execution of all verbal and nonverbal behaviors. Its approach is based on the assumption that continuous speech and gesture are co-produced in successive chunks, each of which being a synchronized pair of one intonation phrase and one co-expressive gesture (as, e.g., the tree in Fig. 8.11), and that the modalities are coordinated at two different levels of an utterance. Within a chunk, the gesture is timed such that its meaning-bearing stroke phase starts before or with the affiliated linguistic elements and spans them. Consequently, the intonation phrase along with its pitch accents is synthesized in advance, using the Festival system for text-to-speech conversion (Black & Taylor, 1997), and timing constraints for coverbal gestural or facial behaviors are drawn from the phoneme time information obtained. Secondly, between two successive chunks, the synchrony between speech and gesture in the forthcoming chunk is anticipated by the movements between the gestures (co-articulation), which may range from the adoption of an intermediate rest position to a direct transition movement. Likewise, the duration of the silent pause between two intonation phrases varies according to the required duration of gesture preparation. Both effects are simulated by our incremental production model, at a time when the next chunk is ready for being uttered (“lurking”) while the former is “subsiding”, i.e., done with executing all mandatory parts (intonation phrase and gesture stroke).

It is at this time, that intra-chunk synchrony is defined and reconciled with the onsets of phonation and movement, and that animations are created that satisfy the movement and timing constraints now determined. NUMACK is able to generate all animations required to drive the skeleton as specified, in real-time and from the scratch. This capability, which is indispensable for the level of generativity targeted here, where the microplanner should be able to come up with novel iconic gestures, is achieved by the MAX module. It allocates the body parts for the gesture, expands symmetry or repetitions constraints, and prepares co-articulation effects when another gesture is about to follow. Following a biologically motivated decomposition of motor control, a final motor planning stage breaks down the control problem—to steer the control variables such that the resulting movement meets all constraints—into sub-problems that get solved by specialized planning modules for the hands, the wrists, and the arms. Their solutions are local motor programs (LMPs) that employ suitable computer animation techniques to control sub-movements, i.e., movement in a limited number of joints and for a limited period of time. To create the whole gesture, the LMPs run concurrently and synchronized in abstract motor control programs, in which they autonomously (de-)activate themselves as well as other LMPs. In result, different, yet coordinated motion generators create different parts of a gesture and automatically create context-dependent gesture transitions.

8.7.3 *Generation Examples*

In our current implementation, NUMACK is able to produce a considerable range of directions, using semantically coordinated language and gesture. Surface realization is real-time in that the time to produce an utterance is typically less than the natural pause between two utterances in dialogue. Together with the lightweight implementation of the microplanner in Prolog, NUMACK is the first system that creates a multimodal utterance from a given set of communicative goals (including all factual and spatial knowledge about the referents) in less than one second on a standard PC. Figure 8.13 demonstrates two example utterances, the one in the left picture was generated from the content plan in Figure 8.7.

8.8 Discussion of Generation Results

As described in the previous sections, the NUMACK system was implemented on the basis of our empirical results, and the NUMACK embodied conversational agent is consequently able to realize direction-giving in quite different ways from that of other ECAs, and other non-embodied dialogue systems. However, “different” does not equal “better” and it is therefore important to assess the fit of the NUMACK system both as a cognitive model of the empirical results described above, and in its role as an autonomous direction-giving system. In the first instance, we ask how similar NUMACK’s performance comes to human direction-giving. In the second instance, we ask how effective NUMACK is in guiding people to their destination, and how NUMACK compares to other direction-giving devices, such as maps. And, in particular how effective are the advances that we made in this system – the generation of direction-giving gestures, and the use of landmarks in route descriptions – to actual human use of the system.

With respect to the first question, concerning NUMACK as a model of human direction-giving, we rely on our own evaluation of NUMACK's performance, and our comparison of NUMACK with the human direction-givers whom we have examined. NUMACK displays some natural hand shapes in describing landmarks, and is certainly capable of a wider variety of direction-giving gestures than previous ECAs that have relied on gesture libraries. However, the ways in which NUMACK is *not* human-like are perhaps more salient than NUMACK's successes. Here we notice that NUMACK tends to give directions all in one go – from point A to point B, without asking the direction-follower if s/he can remember this much information at once. This behavior is striking, and allows us to realize that people must use some heuristic to *chunk* direction-giving into segments. In human direction-giving segments might be separated by explicit requests for feedback (“are you still following me”) or perhaps even followed by a suggestion to ask another passerby (“at that point, you might want to ask somebody else where to go”). The study of route direction chunking, and whether it is based on the speaker's *a priori* beliefs about how long directions should be, or on cues that are emitted by the listener, is therefore one of our topics for future research.

We likewise notice that NUMACK uses the left and right hand interchangeably in pointing out the route, and describing landmarks. Something about this seems unnatural, leading us to think that direction-givers must use one hand or the other time after time as a way of marking cohesion among direction-giving segments. In addition there is something unnatural about the way in which NUMACK uses himself as the origo of his direction-giving, as if he is walking through the scene. Looking at his performance, one is led to think that humans might use their hands to follow an imaginary walker along a route. This too is a topic for future research. In each of these instances, it should be noted that only because we have NUMACK as an instantiation of our theory of gesture and speech in direction-giving, are we even able to evaluate the completeness of the theory, and the places in which we have omitted pertinent analysis of the data.

A second and quite separate topic concerns NUMACK's potential role as a direction-giving kiosk. Is NUMACK more effective than the display of a map? Are NUMACK's directions more effective with gestures, or do the gestures not add much at all? And, are the descriptions of landmarks useful, or are they unnecessary? These three questions lead to 6 different conditions of an experiment to test how people assess the naturalness and effectiveness of direction-giving. Those 6 conditions – an outgrowth of $2 \times 2 \times 2$ (gestures vs. no gestures; landmarks vs. no landmarks; NUMACK vs map only) – are pictured in Figure 8.14. This experiment is currently underway.

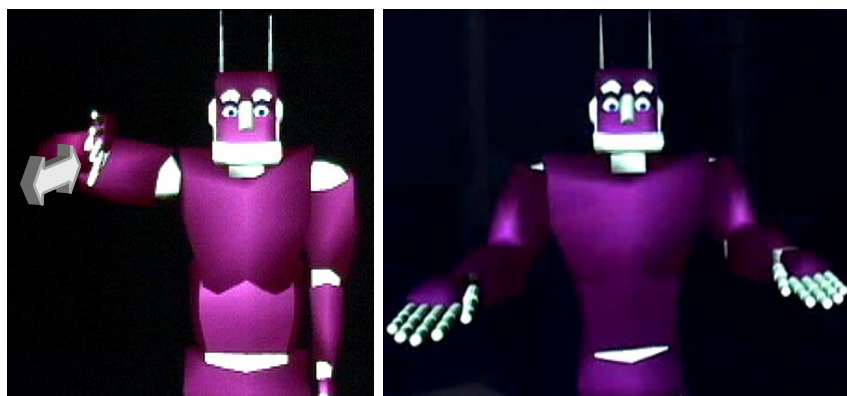


Figure 8.13 NUMACK generation examples: “You will see Cook Hall on your right” (left), and “You will see the Lake ahead” (right).

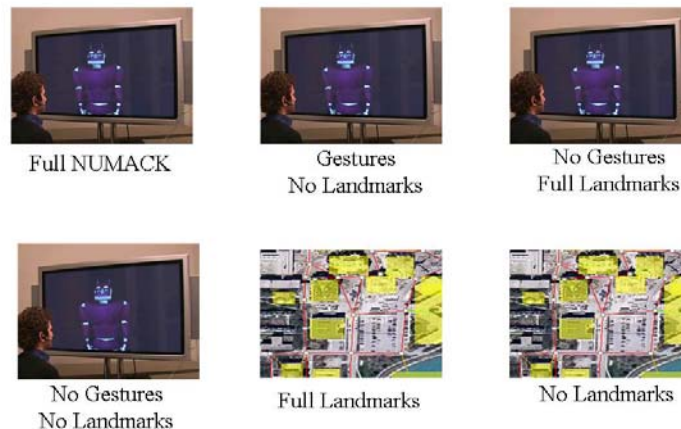


Figure 8.14 Evaluation of NUMACK

8.9 General Discussion and Conclusions

It has been proposed that the different media that participate in direction-giving – diagrams, maps, speech and gesture – all convey one single mental model. And yet, such a claim does not account for the actual form of each medium of expression – why do we use a flat hand with the palm facing inward and the thumb up to represent a path, and a flat hand with the palm facing outward and the palm down to represent walking past a building? And, even more perplexingly, given the lack of standards of form in the gestural medium – even the lack of codified symbols such as those used in maps – how do listeners interpret gesture? In this article we suggest that gesture interpretation and gesture production are facilitated by a layer of meaning that mediates between the morphology of the gesture, and the visuo-spatial attributes of the thing in the world that gesture represents. This level of meaning we call the *image description feature*, and it allows gestures themselves to remain underspecified (lacking consistent form-meaning pairings) and yet meaningful in the context of the speech with which they co-occur.

Analysis of 28 direction-giving sessions allowed us to adduce evidence for the image description feature. An analysis of flat horizontal, flat vertical, and flat sideways handshapes revealed that there was a significant correlation between these features of gesture morphology and similar features in the objects in the world to which these gestures referred in context. In fact, such a result can seem trivial at best – *of course* iconic gesture resembles that which they are iconic of. And yet most researchers have looked for resemblances at the level of the whole gesture, and have not found it (apart from a very small number of culturally specific gestures called *emblems*). And, were the connection so obvious, we would not have found so many instances of false negatives – cases where the visuo-spatial feature was not represented by a similar feature in gesture. Part of the issue, we believe, comes from the simultaneity of gesture – unlike phonemes, for example, units of meaning do not line up neatly one after another in gesture, but occur simultaneously in packages – flat handshape **and** movement away from the body **and** bouncing manner. In addition, IDFs are chosen for their salience *in a particular context*. That is, the same object in the world may very well be conveyed quite differently, depending on what aspect of it is salient in different contexts. Or particular morphological features may be carried over from previous gestures, as part of catchments. Or particular gestures may last longer, and be made more clearly, as a function of the retrievability of the referent. It is clear that we still have a long way to go in order to understand the relationship between gesture and language in particular discourse and pragmatic contexts.

Our empirical results to date nevertheless are strong enough to support the concept of a mediating level of meaning that links gestural features and meaning. And this framework has allowed us to model direction-giving in an embodied conversational agent, and escape the gestionary approach to gesture generation. To this end, we have implemented an integrated, on-the-fly microplanner that derives coordinated surface forms for both modalities from a common representation of context and domain knowledge. In extending the SPUD microplanning approach to gesture planning, lexical entries were replaced with form feature entries; LTAG trees were replaced with feature structures, more closely resembling the global and synthetic nature of gesture; and pragmatic constraints were carried over to guide gesture use in context.

We continue to analyze our empirical data to refine our model, and to find further patterns in the way iconic gesture expresses visual domain knowledge in order to extend the system's generation capabilities. We believe that our approach to microplanning is one step closer to a psychologically realistic model of a central step in utterance formation. However, a range of open questions still need to be investigated, and evaluation of our system will help us shed light on some of these. Such questions include whether a higher degree of interaction might be necessary between the two separate, but interacting planning processes for language and gesture; or whether one unified, qualitative, logic-based representation is sufficient to represent the knowledge required for planning the surface structure of both modalities. In the meantime, we have established a relationship between words and images, language and gesture, where the latter remains underspecified until joined to the former in a particular pragmatic context, and where taken together, words and gestures provide a window onto the representation of space by both humans and humanoids.

Acknowledgements

The authors gratefully thank Matthew Stone for valuable input, Marc Flury and Joseph Jorgenson for technical assistance, and Jens Heemeyer for help with data annotation. This work was supported by a fellowship from the Postdoc-Program of the German Academic Exchange Service (DAAD).

References

- Black, A., & Taylor, P. (1997). *The Festival Speech Synthesis System: System documentation*. Retrieved from.
- Cassell, J., McNeill, D., & McCullough, K.-E. (1999). Speech-Gesture Mismatches: Evidence for One Underlying Representation of Linguistic and Nonlinguistic Information. *Pragmatics and Cognition*, 7(1), 1-33.
- Cassell, J., & Prevost, S. (1996, October). *Distribution of Semantic Features Across Speech and Gesture by Humans and Computers*. Paper presented at the Workshop on the Integration of Gesture in Language and Speech, Newark, DE.
- Cassell, J., Stone, M., & Yan, H. (2000). *Coordination and Context-Dependence in the Generation of Embodied Conversation*. Paper presented at the INLG 2000, Mitzpe Ramon, Israel.
- Cassell, J., Vilhjálmsson, H., & Bickmore, T. (2001). *BEAT: The Behavior Expression Animation Toolkit*. Paper presented at the SIGGRAPH 01, Los Angeles, CA.
- Chi, D., Costa, M., Zhao, L., & Badler, N. (2000). *The EMOTE Model for Effort and Shape*. Paper presented at the SIGGRAPH '00, New Orleans, LA.
- Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.
- Conklin, E. J., & McDonald, D. D. (1982). *Saliency: The key to the selection problem in natural language generation*. Paper presented at the 20th Annual Meeting of the Association for Computational Linguistics, Toronto.
- Couclelis, H. (1996). Verbal directions for way-finding: Space, cognition, and language. In J. Portugali (Ed.), *The construction of cognitive maps* (pp. 133-153). Dordrecht ; Boston: Kluwer Academic Publishers.
- Daniel, M. P., Heiser, J., & Tversky, B. (2003). *Language and diagrams in assembly instructions*. Paper presented at the European Workshop in Imagery and Cognition, Pavia, Italy.
- Denis, M. (1997). The description of routes: A cognitive approach to the production of spatial discourse. *Current Psychology of Cognition*, 16, 409-458.
- Emmorey, K., Tversky, B., & Taylor, H. A. (2001). Using space to describe space: Perspective in speech, sign, and gesture. *Spatial Cognition and Computation*, 00, 1-24.
- Forbus, K. (1983). Qualitative reasoning about space and motion. In D. Gentner & A. L. Stevens (Eds.), *Mental Models* (pp. 53-73). Hillsdale N J: Erlbaum.
- Gao, Y. (2002). *Automatic extraction of spatial location for gesture generation*. MIT, Cambridge, MA.
- Green, N., Carenini, G., Kerpedjiev, S., & Roth, S. F. (1998, August 10-14). *A Media-Independent Content Language for Integrated Text and Graphics Generation*. Paper presented at the Workshop on Content Visualization and Intermedia Representations at COLING and ACL '98, University of Montreal, Quebec.
- Guhe, M., Habel, C., & Tschander, L. (2003). *Incremental production of preverbal messages with inC*. Paper presented at the The Logic of Cognitive Systems. The 5th International Conference on Cognitive Modeling, Bamberg, Germany.

- Heiser, J., Phan, D., Agrawala, M., Tversky, B., & Hanrahan, P. (2004). *Identification and validation of cognitive design principles for automated generation of assembly instructions*. Paper presented at the Advanced Visual Interfaces '04.
- Heiser, J., & Tversky, B. (2003). *Characterizing diagrams produced by individuals and dyads*. Paper presented at the Workshop on Interactive Graphics, London.
- Heiser, J. L., Tversky, B., Agrawala, M., & Hanrahan, P. (2003). *Cognitive design principles for visualizations: Revealing and instantiating*. Paper presented at the Cognitive Science Society Meetings.
- Herskovits, A. (1986a). *Language and spatial cognition*. Cambridge: Cambridge University Press.
- Herskovits, A. (1986b). *Language and spatial cognition : an interdisciplinary study of prepositions in English*. Cambridge: Cambridge University Press.
- Jackendoff, R., & Landau, B. (1991). Spatial Language and Spatial Cognition. In D. J. Napoli, J. A. Kegl & e. al. (Eds.), *Bridges between Psychology and Linguistics: A Swarthmore Festschrift for Lila Gleitman* (pp. 145-169). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Joshi, A. (1987). The relevance of tree adjoining grammar to generation. In G. Kempen (Ed.), *Natural Language Generation: New Results in Artificial and Intelligence, Psychology and Linguistics* (pp. 233-252). Boston: Kluwer Academic Publishers.
- Kerpedjiev, S., Carenini, G., Green, N., Moore, J., & Roth, S. (1998). *Saying It in Graphics: from Intentions to Visualizations*.
- Kopp, S., Tepper, P., & Cassell, J. (2004). *Towards Integrated Microplanning of Language and Iconic Gesture for Multimodal Output*. Paper presented at the International Conference on Multimodal Interfaces (ICMI), State College, PA.
- Kopp, S., & Wachsmuth, I. (2004). Synthesizing Multimodal Utterances for Conversational Agents. *The Journal Computer Animation and Virtual Worlds*, 15(1), 39-52.
- Krauss, R. M., Morrel-Samuels, P., & Colasante, C. (1991). Do Conversational Hand Gestures Communicate? *Journal of Personality and Social Psychology*, 61(5), 743-754.
- Landau, B., & Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral & Brain Sciences*, 16(2), 217-265.
- Lozano, S., & Tversky, B. (submitted). Communicative gestures benefit communicators.
- McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL/London, UK: The University of Chicago Press.
- McNeill, D., & Levy, E. (1982). Conceptual representations in language activity and gesture. *Speech, Place, and Action*, 271-295.
- Milde, J.-T., & Gut, U. (2002). *The TASX-environment: an XML-based toolset for time aligned speech corpora*. Paper presented at the 3rd International Conference on Language Resources and Evaluation (LREC), Las Palmas.
- Nijholt, A., Theune, M., & Heylen, D. (2005). Embodied Language Generation. In O. Stock & M. Zancanaro (Eds.), *Intelligent Information Presentation* (Vol. 27, pp. 47-70): Springer.
- Peirce, C. (1955). *Philosophical Writings of Peirce*. New York: Dover Publications.
- Perlin, K., & Goldberg, A. (1996, August 4-9). *Improv: A System for Scripting Interactive Actors in Virtual Worlds*. Paper presented at the SIGGRAPH '96, New Orleans, LA.
- Poesio, M. (1996). Semantic Ambiguity and Perceived Ambiguity. In K. van Deemter & S. Peters (Eds.), *Ambiguity and Underspecification*. Palo Alto: CSLI.
- Poesio, M. (2005). *Incrementality and Underspecification in Semantic Processing*. Palo Alto: CSLI.
- Prillwitz, S. (1989). *HamNoSys : Hamburg Notation System for Sign Languages : an introductory guide* (ver. 2.0 ed. Vol. 5). Hamburg: Signum Press.
- Reiter, E., & Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge: Cambridge University Press.
- Rickel, J., Marsella, S., Gratch, J., Hill, R., Traum, D., & Swartout, W. (2002). Toward a New Generation of Virtual Humans for Interactive Experiences. *IEEE Intelligent Systems*, 17(4), 32-38.
- Saussure, F. d. (1985). *Cours de Linguistique Générale*. Paris: Payot.
- Sowa, T., & Wachsmuth, I. (2003). *Coverbal Iconic Gestures for Objects Descriptions in Virtual Environments: An Empirical Study*". Paper presented at the Gestures. Meaning and Use, Edicoes Fernando Pessoa.
- Stone, M., Doran, C., Webber, B., Bleam, T., & Palmer, M. (2003). Microplanning with communicative intentions: the SPUD system. *Computational Intelligence*, 19(4), 311-381.

- Talmy, L. (2000). *Toward a cognitive semantics*. Cambridge Mass: MIT Press.
- Talmy, L., University of California Berkeley. Institute of Cognitive Studies., & University of California Berkeley. Cognitive Science Program. (1983). *How language structures space*. Berkeley: Cognitive Science Program Institute of Cognitive Studies University of California at Berkeley.
- Taylor, H. A., & Tversky, B. (1992). Spatial mental models derived from survey and route descriptions. *Journal of Memory and Language*, 31, 261-282.
- Taylor, H. A., & Tversky, B. (1996). Perspective in spatial descriptions. *Journal of Memory and Language*, 35, 371-391.
- Towns, S., Callaway, C., & Lester, J. (1998). *Generating Coordinated Natural Language and 3D Animations for Complex Spatial Explanations*. Paper presented at the AAAI-98.
- Traum, D., & Rickel, J. (2002). *Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds*. Paper presented at the Autonomous Agents and Multi-Agent Systems, Melbourne.
- Yan, H. (2000). *Paired Speech and Gesture Generation in Embodied Conversational Agents*. Unpublished Masters of Science, MIT, Cambridge, MA.
- Young, R. M., & Moore, J. D. (1994). *DPOCL: A Principled Approach to Discourse Planning*. Paper presented at the 7th International Workshop on Natural Language Generation, Kennebunkport, ME.