Producing Semantically Appropriate Gestures in Embodied Language Generation

by

Obed E. Torres
B.S. Computer Engineering
University of Puerto Rico, 1994

SUBMITTED TO THE PROGRAM IN MEDIA ARTS AND SCIENCES,
SCHOOL OF ARCHITECTURE AND PLANNING,
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1998

Signature of Author ____

Program in Media Arts and Sciences
October 17, 1997

Certified by ____

Justine Cassell
AT&T Career Development Professor of Media Arts and Sciences
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by ____

Stephen A. Benton
Chairman, Department Committee on Graduate Students
Program in Media Arts and Sciences

Producing Semantically Appropriate Gestures in Embodied Language Generation

by

Obed E. Torres
Submitted to the Program in Media Arts and Sciences,
School of Architecture and Urban Planning,
on October 17, 1997, in partial fulfillment of the requirements for the degree of
Master of Science

Abstract

In this thesis, I address the problem of producing semantically appropriate gestures in embodied language generation. Most of the previous work on gesture in interactive systems has focused on gestural languages independent of speech content. Research on embodied language generation has directed attention to the generation of spontaneous gestures that accompany speech in face-to-face interactions. Implemented systems for embodied language generation are able to test the placement and realization of gestural phenomena in relation to intonation and information structure. However, important aspects of the problem have not been fully addressed such as generating gestural forms from semantic information and deciding what semantic information should be conveyed across speech and gesture modalities. I extend, in two ways, an existing body of work on the rule-based generation of semantically appropriate gestures. First, by proposing hypotheses about the distribution of semantic information across the speech and gesture modalities and the realization of semantic information in gestural forms. Second, by building a prototype that serves as a vehicle for the exploration of some of the proposed hypotheses and experimentation with the generation of semantically appropriate gestures.

Thesis Supervisor: Justine Cassell

Tittle: AT&T Career Development Professor of Media Arts and Sciences

Producing Semantically Appropriate Gestures in Embodied Language Generation

by

Obed E. Torres

The following people served as readers for this thesis:

Reader ______________________________

Alex Pentland
Toshiba Professor of Media Arts and Sciences
MIT Program in Media Arts and Sciences

Reader ______________________________

Marilyn Walker
Principal Research Staff Member
AT&T Labs Research

Acknowledgments

Writing an acknowledgments section is a self-defeating undertaking. By mentioning certain individuals I will be marking the fact that I do not mention others. That being the case, I will minimize my sense of defeat by strictly acknowledging only those whose contributions are obviously salient because they were in the immediacy of the execution of my thesis work. Without a particular order in mind, I express my gratitude to the following individuals:

To Linda Peterson for suggesting alternatives during periods of apparent impasse.

To Laurie Ward and Teresa Castro for quick assistance with administrative details of thesis work.

To Glen Sherman for diligent responsiveness to my requests of help with thesis mail.

To Justine Cassell for many exposures to her research voice and refined problems representations.

To Marilyn Walker for providing the distinctive comments of a scholarly accurate eye.

To Alex Pentland for important reminders of the relations among audience, fluidity, and lucidness.

To Scott Prevost for innumerable research discussions of unparalleled clarity and acute grace.

To Joey Chang for implementation of the graphical animation engine for the thesis prototype.

To Hannes Vilhjalmsson for uniquely sympathetic, almost infinite, help with the video equipment.

To Mukesh Singh for incommensurably kind aid with the video of the prototype.

To Nick Monfort for inexhaustible patience and competence with the proofreading of this thesis.

To Marina Umaschi for detailed criticism of my thesis proposal.

To Erin Panttaja for extensive comments on my thesis proposal.

To Julia Burch for articulate discussions regarding negotiation and the approval of thesis work.

To Jennifer Glos and Janet Cahn for unclassifiable suggestions at different stages of thesis work.

To David Koons, Joshua Bers, Alberto Castillo, Lee Cambell, Sumit Basu, and Austina De Bonte for immensely useful conversations, although sporadically brief, about the nature of thesis work.

If I have forgotten to mention someone I apologize. Unfortunately, this is another misfortune of singling out individuals. The famous single exists at the expense of the anonymous many. Many relevant things get done by committed and courageous individuals of whom we do not know, like those who walked with Gandhi to the sea or protested with King at lunch counters.

Contents

1. Introduction

1.1 Motivation

When people participate in face-to-face interaction they spontaneously produce gestures while they speak. These gestures enhance or confirm the information conveyed by words. They occur at the same time as speech and carry a meaning associated with the content of speech. This suggests the existence of an intimate link between the production of gesture and speech. Research shows that people produce the majority of gestures during speech articulation (McNeill 1992). These gestures expand and support the content of the accompanying speech (Kendon 1972, McNeill 1992). They can also reinforce speech such as when there is alignment between the most effortful parts of the gestures and intonational prominence (Kendon 1980). Several researchers have argued that gesture and speech are different communicative expressions of a single mental representation (Kendon 1972, McNeill 1992, McNeill 1996, Cassell et. al. in press).

Gesture and speech do not always convey the same information about the meaning of a communicative intention. Gesture might depict aspects of meaning that were not described in speech. Kendon has suggested that what is difficult to express in speech may be conveyed by gesture (Cassell et. 1994b, Kendon 1994). Visual information that is elusive to speech or dynamic scene features difficult to convey with speech, such as the simultaneity of two events or the spatial locations of objects, may be expressed by gesture (Cassell et. al. 1994b, Kendon 1994). McNeill and Levy show that gesture and speech convey different narrative information (McNeill & Levy 1982). Cassell and McNeill discuss how gesture often represents the point of view of the speaker when this is not present in the context of the speech (Cassell & McNeill 1991). None of these accounts of the complementarity or redundancy of information across modalities identify how to determine the distribution of communicative load. That is, none of them explains how and what information is distributed across speech and gesture. A more comprehensive account of the semantic complementarity and redundancy among modalities has important implications for the study of representational processes and the nature of language, and issues such as investigating whether visual ideas are encoded as spatial structures or if representations are modality-specific

(Kendon 1986).

Recently, Cassell and Prevost hypothesized rules to predict when redundant and complementary information are conveyed in speech and gesture (Cassell & Prevost 1996, Cassell & Prevost under review). Although knowing when to convey redundant or complementary information is a step toward a more comprehensive account, several aspects remain to be addressed, such as determining what content is conveyed in gesture and in which gestural form, and determining what content is realized in speech, what is realized in gesture, and how the distribution may be represented.

1.2 Contribution

The generation of paralinguistic behaviors in conjunction with the generation of language is termed embodied language generation. Research on embodied language generation has contributed to the study of the relation of speech and gesture (Cassell et. al. 1994b, Cassell & Prevost under review). Implemented systems for embodied language generation are able to test the placement and realization of gestural phenomena in relation to intonation and information structure (Cassell et. al. 1994b). However, there are other aspects of the relation between speech and gesture that a more comprehensive account of the interplay of speech and gesture should address. If, as has been suggested by several researchers, speech and gesture arise from a single underlying representation, then a more comprehensive account of the semantic relation between speech and gesture should explain the expressive demands of that representation; which features are expressed in speech and which in gesture; and how they are expressed in speech and gesture (Kendon 1972, McNeill 1992, McNeill 1996, Cassell et. al. in press). This thesis addresses these research questions within the context of producing semantically appropriate gestures in embodied language generation. In particular, it addresses what content is conveyed in gesture and in which gestural forms. Also this thesis addresses what is realized in gesture and speech, and how the distribution may be represented.

An evaluation of the results of the thesis would entail conducting memory tasks and comparisons, as has been done in previous work on speech-gesture mismatch experiments, with subjects that are exposed to narrations containing instances of complementary and redundant information and determining to what extent they rely as much on the information conveyed by the gesture as that conveyed by speech and incorporate both in the understanding of what they heard (Cassell et. al. in press). Although such evaluation is beyond the scope of this thesis, the implemented prototype facilitates conducting it.

This work differs in at least two ways from trends in existing work on gesture as documented in research areas such as computer vision, human-computer interfaces, linguistics, and cognitive science. First, most of the previous work on gesture in interactive systems focuses on gestural languages independent of the content of speech (Cassell in press). In this thesis, the research work concentrates on spontaneous gestures that accompany speech. Second, most of the research in cognitive science or linguistics addresses the interaction of speech and gestures in a descriptive manner (Cassell & Prevost under review). Descriptive approaches do not account for the underlying processes of observable patterns. None of the accounts on the semantic relation between speech and gesture identify how to determine the distribution of communicative load. In this thesis, the research combines theoretical and practical approaches. The theoretical approaches include deriving hypotheses from the literature on the semantic relation between speech and gesture, semantic structures, and the relation between spatial language and spatial cognition. The practical approaches include analyzing experimental data using formalisms of semantic structures and discussing empirical instances that do or do not support the hypotheses. The practical approaches also include building a prototype that uses some of the derived hypotheses as heuristics for the generation of semantically appropriate gestures in embodied language generation.

1.3 Overview

This chapter has provided an introduction to the research work of this thesis by presenting the motivations and reasons to pursue this research, its overall contributions, and what evaluating its results would entail. The rest of the thesis is divided into six chapters that, as a whole, contextualize the problem, present the methodological approaches, discuss relevant results, and describe applications of the findings in the construction of a prototype. Chapter 2 reviews related work on the communicative function of gestures, summarizes trends in the work on gesture in interactive systems, and comments on work in gesture within the area of embodied language generation.

Chapter 3 describes the experimental data, transcription procedures, and methodological approaches that were used to analyze the semantic relation between speech and gesture in a sample of experimental data. Chapter 4 discusses relevant results of the empirical study to advance hypotheses about the distribution of semantic information across the speech and gesture modalities. Chapter 5 discusses relevant results of the empirical study to advance hypotheses about the realization of semantic information in gestural forms. Chapter 6 describes a prototype built to explore and experiment with the generation of semantically appropriate gestures. Chapter 7 summarizes and evaluates the main contributions of this research and future work to extend it.

2. Context of the Work

This chapter summarizes relevant previous work on gesture in theoretical and practical research areas. This previous work is related to or supports research on the semantic relation between speech and gesture.

2.1. Gesture and Communication

People spontaneously produce gestures while speaking to expand and support accompanying words (Kendon 1972, McNeill 1992). This elaboration and enhancement of the content of speech provides an index to the underlying thematic organization of the discourse and point of view in events (McNeill 1992). Gesture can also provide an index to underlying reasoning processes of problem solving that are not present in the content of speech (Church & Goldin-Meadow 1986, Goldin-Meadow et. al. 1993).

Gestures are sometimes produced in the absence of a listener although more gestures are produced when a hearer is present (Cassell in press, Cohen & Harrison 1973, Cohen 1977, Rime & Sciaraturea 1991). Although effective communication takes place in the absence of gestures, such as in phone conversations, it has been demonstrated that when speech is ambiguous, listeners utilize gestures to interpret it (Cassell in press, Short et. al. 1976, Williams 1977, Rogers 1978, Thompson & Massaro 1986). Also, when subjects were presented with slightly different information in speech and gestures that added to or contradicted each other, they relied as much on the information conveyed by gestures as that conveyed by speech, and incorporated both to build a single representation of the information conveyed in the two modalities (Cassell et. al. in press).

Gestures generally co-occur with their semantically parallel linguistic units, except in the cases of pauses or syntactically complex speech, in which case the gesture occurs first (McNeill 1992). There is also synchronization between individual gestures and words, so that the most effortful part of the gesture occurs with or just before the most intonationally prominent syllable of the accompanying speech (Kendon 1972, McNeill 1992, Tuite 1993).

Ekman and Friesen identified five categories of nonverbal behaviors with semantic components (Ekman and Friesen 1969). What they termed as illustrators, those nonverbal behaviors interpreted as being part of the speech performance, have been the object of further study mainly by Kendon and McNeill in their categorizations of gesture in relation to the content of speech (Kendon 1980, McNeill 1992).

According to McNeill, 90% of all gestures occur within the context of speech (McNeill 1992). These spontaneous gestures are of four types: iconic, metaphoric, deictic, and beat. A fundamental assumption underlying the division of these categories is that the speaker is mapping an internal representation into patterned hand motor articulations. This assumption is grounded on the observation that in many instances the form of the gesture resembles features of actual objects and actions.

Iconic gestures describe features of an action or event. The form of the gesture may depict the manner in which an action was carried out or the viewpoint from which an action is narrated. An example of this type of gesture from an observer point of view is the use of a hand, with extended fingers, moving toward a fixed hand, with a flat palm, until they strike each other while the speaker is saying “The CAR HIT the wall.”

Metaphoric gestures describe features of a concrete object to represent an abstract concept. The form of the gesture is related to a culturally dependent common metaphor for a concept that has no physical form. The conduit methaphoric is a common metaphoric gesture. This metaphoric gesture objectifies the information of the accompanying speech through a representation of a concrete object that can be held and manipulated with the hands. An example of this type of gesture is the creation of a concrete object to represent the concept of a scene and the following narration of it by the rise of both hands, as if holding up an object as if it were a container, and the movement of both hands away from the body, as if opening the object like a container, while saying “It WAS A CAR ACCIDENT scene in Elm St.”

Deictic gestures locate in space those discourse entities with or without a physical instantiation. They occur when discourse entities are introduced or continuously referred to. The form of the gesture is a pointing index finger or a whole hand representing the discourse entity in space. An example of this type of gesture is an extended index finger pointing to the corner sidewalk while saying “Did the car accident occur at THAT corner or at the exit of the parking lot?”

Beat gestures are abstract visual indicators to emphasize discourse-oriented functions. They are small movements of constant form throughout the content of the accompanying speech. They occur with speech repairs or to emphasize a linguistic item in relation to others. An example of this type of gesture is a quick left hand waving while saying “Please, CONTINUE with your description of the scene.”

Most of the research on gesture and communication within cognitive science or linguistics addresses the interaction of speech and gestures in a descriptive manner (Cassell & Prevost under review). Descriptive approaches do not account for the underlying processes of observable patterns. None of the accounts on the semantic relation between speech and gesture identify how to determine the distribution of communicative load. Research addressing the question of whether gestures are always redundant has dealt with the role of gesture in expressing concepts that are difficult to express in language (Cassell & Prevost 1996, McNeill 1996). Such research has offered insights about certain semantic features expressed in gesture and not in speech, or expressed in both. Still this research lacks specificity and consistency because the analyses do not use semantic features derived from a formalism of semantic primitives and their principles of combination.

2.2. Gesture and Interactive Systems

Most of the previous work on gesture in interactive systems has focused on gesture as a language rather than gesture as part of a communicative action (Cassell in press). The finitude of gestural languages makes them suitable to template-based matching. Murakami and Taguchi implemented a gesture recognizer of alphabet gestures in the Japanese Sign Language (Murakami & Taguchi 1991). Vaananen and Bohm also built a system using a neural network recognizer to map specific hand configurations to system commands (Vaananen & Bohm 1993). Vision-based gesture recognition research has made possible the implementation of systems that recognize small libraries of emblematic gestures and structured hand gestures such as those found in the American Sign Language (ASL) (Queck 1994, Starner & Pentland 1995).

Research on user interfaces has concentrated on gestures as substitutes for words, including multi-modal systems that combined speech and demonstrative gesture. The Put-That-There system used speech recognition and a space sensing device to integrate the speech of the user with the position of the cursor to resolve references to items in a wall-sized screen (Bolt 1980). In another system, speech and gestures were integrated to allow manipulation of graphical objects (Bolt & Herranz 1987). Koons built a multimodal interpreter of speech, gestures, and eye movements to resolve deictic references to objects in a wall-sized screen (Koons et. al. 1993). In this system the speech modality drives the analysis of the integrated information. That is, if information is missing from the content of speech, the interpreter searched for it in the other modalities (Koons 1994). Bers integrated the speech and pantomimic gestures of the user to direct animated creature behaviors such as the manner in which a bee moves its wings (Bers 1995).

2.3. Gesture and Embodied Language Generation

Research on autonomous agents, human-computer interfaces and virtual environments has stirred interest in embodied language generation. Traditionally, natural language generation has focused on the production of written text. Face to face interaction includes spoken language (words with contextually appropriate intonation), facial expressions (lip shapes, head motions, gaze behavior) and hand gestures (semantically appropriate hand shapes and trajectories). Research on embodied language generation is concerned with several research problems related to the intersection of verbal and nonverbal behaviors in face-to-face interaction, including facial expressions, intonation, and gesture (Cassell et. al. 1994a, Pelachaud et. al. 1996, Prevost 1996, Torres et. al. 1997). The research area of embodied language generation encompasses research issues such as graphical animation of human-like figures with faces and hands and spoken language generation with appropriate intonation, gestures, and facial expressions (Cassell & Prevost under review).
User testing of embodied human-like agents that exhibit face-to-face conversational behaviors

has shown that the presence of nonverbal feedback behaviors increases believability and effectiveness in the interaction (Thorisson 1996).

Animated Conversations was the first computational implementation of human-like figures conveying believable and contextually appropriate gestures, facial expressions, and spoken intonation (Cassell et. al. 1994a). The domain of the dialogue was banking transactions. The scenario was a bank teller in a transactional conversation with a customer wanting to withdraw money. The implemented system used the timing of the intonation to determine the placement of gesture (Cassell et. al. 1994b). The system generated the distribution of gestures using information structure. Information structure describes the relation between the content of the utterance and the emerging discourse context. Gestures were generated along words or phrases that were marked as rhematic contributions to the discourse. The system generated the occurrence of the different types of gestures through a one-to-one association of a gestural form with a concept. Although this system showed that it was possible to predict and generate speech and gesture from a common underlying semantic representation, the gestures that were generated were redundant. The system was able to place gestures wherever they were expected as a function of the discourse structure and according to the appropriate temporal relations between speech and gesture. However, there was no basis for the distribution of the communicative load. The system provided an inadequate solution, a gesture dictionary that provided particular gesture forms for particular concepts, only allowing the generation of predefined constant gestural forms for each instantiation of the concepts.

In their presentation of a framework for the generation of speech and gesture, Cassell and Prevost discuss a preliminary analysis of experimental data in which subjects were shown a segment of a Road Runner cartoon (Cassell & Prevost 1996, Cassell & Prevost under review). Each subject told the story to a naive subject. A partial analysis of this experimental data allowed the formulation of several heuristics on the basis of information structure distinctions to predict when redundant and complementary information occurs. Cassell and Prevost used “theme” and “rheme” to specify the information structural components of an utterance (Halliday 1967). The thematic part of an utterance represents the chunk of information that links the utterance to the previous discourse and specifies what the utterance is about. The rhematic part of an utterance represents the chunk of information that specifies what is contributed with respect to the theme, that is, what is new or interesting about the theme. Within theme and rheme there are other information structural categories, among them, the focus. The focus of thematic material marks contrastive items in the theme. The focus of a rhematic material marks contrastive or new items in the rheme. New items are those not previously mentioned or salient in the discourse. Contrastive items are those in direct contrast with other salient items in the discourse.

Using the information structure distinctions of theme, rheme, focus, newness, and contrastiveness, Cassell and Prevost hypothesized three rules, as presented in (1), (2) and (3), to predict when redundant and complementary information are conveyed across speech and gesture.

(1) The new rhematic information rule is that rhematic information with a focus marking newness indicates complementary information across both modalities.

(2) The contrastive rhematic information rule is that rhematic information with a focus marking contrastiveness indicates redundant information realized in both modalities.

(3) The contrastive thematic information rule is that thematic information with a focus marking contrastiveness indicates redundant information expressed by both modalities.

This set of hypotheses, by Cassell and Prevost, about when to convey redundant or complementary information, is a step toward a more comprehensive account on the semantic relation of speech and gesture. Still, several aspects remain to be addressed, such as determining what content is conveyed in gesture, and determining what content is realized in speech, what is realized in gesture, and how the distribution may be represented.

If, as has been suggested by several researchers, speech and gesture arise from a single underlying representation, then a more comprehensive account of the semantic relation between speech and gesture should explain the expressive demands of that representation; which features are expressed in speech and which in gesture; and how they are expressed in speech and gesture (Kendon 1972, McNeill 1992, McNeill 1996, Cassell et. al. in press 1994a). This thesis addresses these research questions within the context of producing semantically appropriate gestures in embodied language generation. In particular, it addresses what content is conveyed in gesture and in which gestural forms. Also this thesis addresses what is realized in gesture and speech, and how this distribution is represented. In this thesis, the research combines theoretical and practical approaches. The theoretical approaches include deriving hypotheses from the literature on the semantic relation between speech and gesture, semantic structures, and the relation between spatial language and spatial cognition. The practical approaches include analyzing experimental data using formalisms of semantic structures and discussing empirical instances that support or contradict the hypotheses. The practical approaches also include building a prototype that uses some of the derived hypotheses as heuristics for the generation of semantically appropriate gestures in embodied language generation.

2.4. Semantics Theories and Parallel Semantic Fields

Componential, procedural, frame, and lexical semantics are among the most well-known semantic theories. Componential semantics defines the sense of a word in terms of a set of features that distinguish it from other words in the language (Katz & Fodor 1960). In componential semantics, the meaning of a lexical item is analyzed in terms of terminal elements of feature decompositions. Procedural semantics defines the meaning of words in terms of a set of instructions related to the usage of the word (Miller & Johnson-Laird 1976). In procedural semantics, the meaning of a lexical item is analyzed in terms of procedure extractions in the process of lexical recognition or production. Frame semantics defines the meaning of a word with reference to a structured background of experience, belief, and practices that motivate the concept encoded by the word (Fillmore 1982). In frame semantics, the meaning of a lexical item is analyzed in terms of background frames of meaning structures associated to certain lexical and syntactic patterns. Lexical semantics defines the meaning of a word in terms of representations of the world and its relation to language (Saint & Viegas 1995). In lexical semantics, the meaning of a lexical item is analyzed in terms of the forms of semantic representations and the relations between semantic and syntactic levels of representation. Work on lexical semantics by Talmy and Jackendoff has provided schemes for analyzing concepts of spatial location and motion (Talmy 1985, Jackendoff 1990). Talmy's work provides important analyses of spatial concepts in relation to other semantic components and cross-language grammars. Jackendoff's conceptual semantics establishes a formal correspondence between external language, seen as an artifact, and internal language, seen as a body of internally encoded information. The next chapter briefly describes the analytical schemes provided by Talmy and Jackendoff.

The work of Talmy and Jackendoff stands out when one considers the claim that when schemes for analyzing concepts of spatial location and motion are appropriately abstracted they can be generalized to parallel semantic fields (Jackendoff 1990, Gruber 1965). This claim is based on the fact that many verbs and prepositions occur in two or more semantic fields and form intuitively related patterns (Jackendoff 1990). Jackendoff illustrates this with four semantic fields: spatial location and motion, possession, ascription of properties, and scheduling (Jackendoff 1990). As presented in (4), (5), (6) and (7), each of these semantic fields contains the verbs "go" or "change," "be," and "keep." The sentences with the "go" verb express change. The sentences with the "be" verb express terminal states. The sentences with the "keep" verb express the causation of a state that extends over a period of time. Similar elements vary the place of the entity. In the spatial field, the entity is located spatially. In the possessional field, the entity belongs to someone. In the ascriptional field, the entity has a property. For the scheduling field, the event is located in a time period.

(4) Spatial location and motion

a. The car went from the bank to the bar.

b. The car is in the parking lot.

c. John kept the car in the garage.

(5) Possession

a. The prize went to Rick.

b. The award is Tod's.

c. Mitchell kept the scholarship.

(6) Ascription of properties

a. John went from ecstatic to bored.

b. John is happy.

c. John kept his mood.

(7) Scheduling of activities

a. The gathering was changed from Monday to Wednesday.

b. The conference is on Friday.

c. Let’s keep the appointement on Thursday.

3. Analyzing the Semantic Relation Between Speech and Gesture

If speech and gesture arise from a single underlying representation, then a more comprehensive account of the semantic relation between speech and gesture should explain the expressive demands of that representation; which features are expressed in speech and which in gesture; and how they are expressed in speech and gesture (Kendon 1972, McNeill 1992, McNeill 1996, Cassell et. al. in press). This thesis addresses these research questions within the context of producing semantically appropriate gestures in embodied language generation. In particular, it addresses what content is conveyed in gesture and in which gestural forms. It also addresses what content is realized in gesture and what is realized in speech, and how the distribution is represented. In this thesis, the research combines theoretical and practical approaches. The theoretical approaches include deriving hypotheses from the literature on the semantic relation between speech and gesture, semantic structures, and the relation between spatial language and spatial cognition. The practical approaches include analyzing experimental data using formalisms of semantic structures and discussing empirical instances that support or contradict the hypotheses. The practical approaches also include building a prototype that uses some of the derived hypotheses as heuristics for the generation of semantically appropriate gestures in embodied language generation.

This chapter explains the systematic approach used in this research to analyze the semantic relation between speech and gesture. The first section describes the sample of experimental data that was transcribed. The second section describes the procedures employed to transcribe the data. The third section describes the methodological approach used to analyze the transcriptions.

3.1. Experimental Data

The data used in this empirical study is a sample of the data collected during a preliminary experiment to examine associations between manner of motion verbs and gesture (Cassell & Prevost

1996). All of it was recorded during narrative discourse. The speaker was brought into the laboratory area for video screening and shown segments of a Road Runner cartoon. After the screening, the speaker was brought into a video room to recount the story to a naive listener. The listener was allowed to interrupt the narrator to ask questions whenever the narration was unclear. This narration was videotaped. The viewing time of the cartoon segment was about five minutes and it usually took about that amount of time or longer to narrate it. The cartoon stimulus was an animated color cartoon of three segments with the same two characters, Wiley Coyote and Road Runner, and the same structure, Wiley Coyote trying to harm Road Runner, but with different entertaining variations. The cartoon segments were selected because of their simple and repetitive content of scenes of motion events, their highly visual display of events in the absence of dialogue, and their likelihood to evoke a large number of iconic gestures.

All participants were given the same instructions. The speaker was told that since the listener does not have prior knowledge of the stimulus and did not watch the cartoon segment it was important to present the story in as detailed and clear a fashion as possible so that the listener could retell it later. Participants were told that the data collection was part of a research effort to study face-to-face interactions without mentioning that gestures were the primary area of interest. All of them consented to be videotaped. All subjects, three of them male and three of them female, were adult native speakers of American English.

The narrations were videotaped using two cameras and a microphone placed so that the upper-body space of the speaker and the listener were completely visible and their voices could be comfortably heard. The narrations were videotaped without altering the focus, zooming in or out, or increasing or decreasing the level of sound. The cameras and the microphone were set up in full view of the participants. The video camera, the microphone, and the video tape were running before participants started their narrations and were not stopped until the narrations ended. The narrator and the listener were placed face-to-face in comfortable chairs with ample free space in front of their upper body so they could move their hands without obstructions.

The experimental sample consists of fifty utterances selected because they contain visible iconic gestures. The fifty utterances consist of approximately ten utterances from five different speakers in three different narrations and two retellings. In all selected instances the hand movements were depicting imagery. Because they were iconic, the gesture bears a close relation to the semantic content of speech. All of them, in their form and manner of execution, depicted aspects of motion scenes. In order to select only iconic gestures, it was necessary to compare the hand movement with the scene described by the speech and then determine if the gesture was exhibiting an aspect of it.

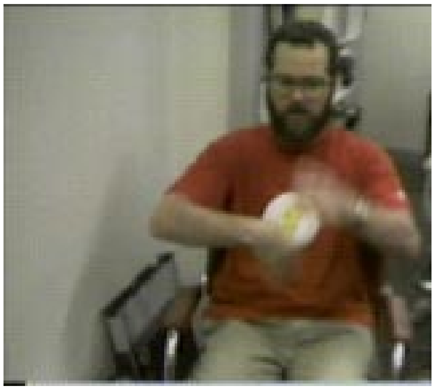


Fig. 1 “he unravels”

3.2. Transcription Procedures

The procedures for transcribing the samples of iconic gestures are based on McNeill’s coding scheme (McNeill 1992). The main coding procedures facilitate transcribing experimental data about the hands, motion, and meaning. For the hands, the following aspects were coded: which hands were used, shape of the hand, and palm and finger orientations. For the motion, the following aspects were coded: shape of trajectory and space where motion takes place. For the meaning of the hands, the following aspects were coded: what they represent and their viewpoint. For the meaning of motion, the following aspects were coded: what it represents and what features are represented, such as manner, direction, and kind of path. Appendix D contains a sample of these transcriptions. Figure 1 shows a video frame of a subject while uttering “he unravels.”

The slow motion capabilities of the video player, the numbering of the video frames and an oscilloscope trace of the speech were used to align gesture and speech. Time stamps were placed at the beginning and at the end of each gestural phrase. The beginning of a gestural phrase is at the onset of the hand movement. The end of a gestural phrase is at the cessation of the hand movement. Speech accompanying the gestural phrases was fully transcribed. Brackets were placed within the gestural phrase to specify the boundaries of the occurrence of the stroke. Kinetically, the stroke is the most effortful part of the hand movement. Semantically, the stroke is the part of the hand movement that bears content.

For transcription purposes, spatial positioning of the gestures was specified with reference to specific points in the gesture space (shoulders, chest, and waist) and zone divisions. The three dimensional gesture space can be divided into five zones covering horizontal and vertical axes (left, right, upper, lower, center) and two more other zones covering the depth axis (away from the body and toward the body). Most of the hand shape specifications were approximations to ASL hand shapes (e.g. A, C, 5, O, and G) that are likely to occur in all narratives (McNeill 1992). If a hand shape did not resemble any of commonly present ASL hand shapes then it was specified in terms of a group of features such as palm configuration (flat, open, or close), which fingers were extended, thumb position (in, out, or over) and finger contacts with each other.

3.3. Methodological Approach

The main methodological approach consisted of a systematic analysis to explore relations between meaning and form expression. The units of the analysis were basic elements in the domain of meaning (spatial concepts of motion and location in motion events) and the domain of form expression (gestural forms and lexical items). The domain of meaning entailed examining semantic features associated with sketching motion events using schemes for analyzing from Jackendoff's and Talmy's work on lexical semantics (Talmy 1985, Jackendoff 1990). For gestures, examination of the domain of expression involved analyzing the forms of the gestures as instantiated in hand shapes and hand trajectories. For lexical items, examination of the domain of expression involved analyzing motion verbs and the prepositions associated with them.

The examination of the correspondence between meaning and form elements is not a one-to-one mapping problem. When looking for patterns and associations between meaning and form elements, the observation of what form elements appear to be expressed when holding a constant domain of meaning is as relevant as the observation of what meaning features are expressed when holding a constant form element. This approach to exploring relations between meaning and form expression has been effectively employed to conduct crosslinguistic characterization of lexicalization patterns and in tackling the computationally complex task of lexical choice for language generation (Talmy 1985, Elhadad et. al. in press).

Talmy’s work on the semantic patterns of motion events offers a scheme of analysis to examine spatial concepts of motion and location (Talmy 1985). It provides important analyses of spatial concepts in relation to other semantic components and cross-language grammars. According to Talmy, a motion event is a situation in which there is movement and/or the maintenance of a stationary location. The “Figure” is the moving or located object with respect to a fixed reference object termed the “Ground.” The “Path” is the course followed or site occupied by the Figure with respect to the Ground. For example in a sentence like “The car turned left around the corner,” the “car” corresponds to the Figure, the “corner” is the Ground and “left” and “around” express Path.

Conceptual semantics assumes that meanings are mentally encoded and decomposable which means that meanings have an internal structure built up from an innate framework of primitives and principles of combination (Jackendoff 1987). Jackendoff’s conceptual semantics establish a formal correspondence between external language, seen as an artifact, and internal language, seen as a body of internally encoded information (Jackendoff 1990). It develops a formal treatment of semantic intuitions. It accounts for the distinctions of meaning and semantic relations among words. It also accounts for the relation between the formal treatment of meaning and the structure of syntax. It situates its rigorous formalisms in an overall psychological framework that integrates linguistic theory with theories of perception and cognition.

Jackendoff’s conceptual semantics argues for the existence of essential units of conceptual structures as conceptual constituents (Jackendoff 1990). These conceptual constituents belong to ontological categories such as THING, EVENT, STATE, PLACE, and PATH. Major syntactic parts of a sentence map into conceptual constituents. Not every conceptual constituent correspond to a syntactic constituent because many conceptual constituents are contained within lexical items.

For example, the syntactic part “the car” corresponds to the conceptual constituent THING, “next to the three” corresponds to PLACE, “across the road” corresponds to PATH, “the car is next to the stop sign” corresponds to STATE, and “the car hit a tree” corresponds to EVENT. The conceptual categories of PLACE, PATH, EVENT and STATE are defined in terms of function-argument structure types as illustrated in (1), (2), (3), and (4).

```
(1)  ( PLACE ( PLACE-FUNCTION ( THING ) ) )

(2)  ( PATH (  PATH-FUNCTION ( THING ) ) )
     ( PATH (  PATH-FUNCTION ( PLACE ) ) )

(3)  ( EVENT ( GO ( ( THING )
                    ( PATH ) ) ) )
     ( EVENT ( STAY ( ( THING )
                      ( PLACE ) ) ) )
     ( EVENT ( CAUSE ( ( THING )
                       ( EVENT ) ) ) )
     ( EVENT ( CAUSE ( ( EVENT )
                       ( EVENT ) ) ) )

(4)  ( STATE ( BE (  ( THING )
                     ( PLACE ) ) ) )
     ( STATE ( EXT ( ( THING )
                     ( PATH ) ) ) )
     ( STATE ( ORIENT (  ( THING )
                         ( PATH ) ) ) )
```

The PLACE function denotes a spatial location in terms of a spatial reference point which belongs to the THING ontological category. PATH functions denote trajectories relative to a THING or a PLACE. Functions such as BE, ORIENT and EXT define the STATE conceptual category. BE denotes a THING placed relative to a PLACE. ORIENT denotes a THING oriented relative to a

PATH. EXT places a THING along a PATH. Functions such as GO, STAY and CAUSE define the EVENT conceptual category. GO denotes the motion of a THING along a PATH. STAY describes the location of a THING at a place for a extended period of time. CAUSE denotes the relation between an EVENT and the THING or EVENT that caused it.

Some of the hypotheses in the next two chapters are motivated with references to Jackendoff's formalisms. Although Talmy's work is also concerned with spatial concepts of motion and location and provides important analyses of spatial concepts in relation to other semantic components and cross-language grammars, it lacks the formal treatment to render a testable theory. Also, in Talmy's work, the articulation of relations between semantic and perceptual components is tacit. On the contrary, Jackendoff's work on conceptual semantics provides general primitives and rigorous principles of combination to form representations of the meaning of linguistic expressions (Jackendoff 1990). It makes explicit its relations to relevant results in the studies of perception and provides a rigorous formalism of primitive decompositions that forces analyses to be concrete and renders a testable theory (Jackendoff 1987, Jackendoff 1990).

4. Distribution of Semantic Information Across Speech and Gesture

Chapter 4 and 5 present hypotheses and discussions of instances in the sample of experimental data support or do not support the proposed hypotheses. There are a few overall remarks about the nature of these discussions. First, generality and context independence are important characteristics of scientific explanations. However, at early stages of the analysis of a problem, the further the analysis is removed from specific and situated discussions the greater the risk of falling into a distorted understanding of the function of things. This could make the analysis seem somewhat premature, dubious, and suspicious. However, this does not have to be the case if, throughout the process, the analysis has included asking meaningful empirical questions and offering sensible justifications to support the choices among alternative claims. Second, despite the observable regularities supporting the proposed hypotheses there are instances in the sample of experimental data that fail to conform to explanations derived from them. Four general explanations could be offered. First, some of these instances could have been performance errors. Second, it could be the case that some of the proposed hypotheses are optional instead of essential, or even special cases of more general hypotheses. Third, there is no comprehensive account of the semantic relation between speech and gesture so there could be competing rules with hierarchies and precedences not well understood yet. Fourth, some of the proposed hypotheses could be interrelated to rules that apply to different aspects of the semantic relation between speech and gesture such as information structure distinctions.

In this chapter I offer three hypotheses, as presented in (1), (2), and (3), regarding the distribution of semantic information across speech and gesture.

(1) The implicit arguments hypothesis is that an implicit argument in a conceptual structure may serve as a representation of what is distributed in gesture as complementary information.

(2) The natural actions hypothesis is that a spatial structure encoding linked to a conceptual structure of a natural action may serve as a representation of what is distributed in gesture as comple-

mentary information.

(3) The marked features hypothesis is that a marked feature in a conceptual structure may serve as a representation of what is distributed in gesture as redundant information.

4.1. The Implicit Arguments Hypothesis

Goldin-Meadow et. al. suggests that gesture may reflect knowledge that is implicit in speech (Goldin-Meadow et. al. 1993). By appealing to this notion, I claim that an implicit argument in a conceptual structure serves as a representation of what is distributed in gesture as complementary information. As briefly discussed in the previous chapter, Jackendoff's work on conceptual semantics provides function-argument structures to encode conceptual representations of meaning. In these structures, implicit meaning is encoded as an empty argument. An empty argument may serve as a representation of what is distributed in gesture as complementary information. That is, implicit information is conveyed as complementary meaning in the gesture modality.

The utterance "the car approaches" illustrates the implicit argument hypothesis. The sentence means that the car traversed an unspecified trajectory. It is implicit that the car traversed a path. Its conceptual representation is of the function-argument structure of type (4) which after argument substitution takes the form of (5). If a gesture accompanies this sentence, it is likely to include a hand, as a thing in motion, following a trajectory in the gesture space, a conceptual space to demonstrate the path, representing the thing in motion traversing a region. Among the types of information, not present in speech and related to the missing argument, that gestures of the speaker may convey are: the type of movement elongation axes that are specified, such as left-right, up-down, or front-back, axis and the type of spatial relations that are imposed whenever the narrator takes the position of observer or character, such as front (being what is facing the speaker) or behind (being what is not facing the speaker).

```
(4)  ( GO ( ( THING )
            ( PATH ) ) )
```

(5) (GO ((car)
(void)))

4.2. The Natural Actions Hypothesis

Jackendoff suggests that natural actions, like natural kinds, are learned more by ostension (“This is what it looks like”), exemplifying their appearance, than by definition (Jackendoff 1987). By appealing to this notion, I claim that a spatial structure encoding linked to a conceptual structure of a natural action serves as a representation of what is distributed in gesture as complementary information. Jackendoff’s conceptual structures encode meaning in an algebraic format. There are types of information, such as shape of objects, that are not easily decomposable into algebraic features because even if such features are identified it is difficult to point out or demonstrate that they are primitives in a system of conceptual representations instead of ad-hoc choices. This type of information has been termed natural kinds (Jackendoff & Landau 1991). Distinctions of meaning among natural kinds are better represented in a terms of spatial structures than in terms of conceptual structures. According to Peterson there is a class of natural actions analogous to natural kinds. Jackendoff proposes that conceptual structures for natural actions such as crawling, dancing, and squirming, are linked to a spatial structure encoding of categories of actions (Jackendoff 1990, Peterson 1987). Jackendoff defines a set of correspondence rules between conceptual structures and the spatial structure encoding as an extension of Marr’s visual representation model (Jackendoff 1987, Marr 1982). The spatial structure encoding is a 3D model representation where the primitive units are coordinate axes and means for generating simple shapes around them. The Actions of moving figures are represented as sequences of motions of the figure parts of the 3D model (Jackendoff 1990, Marr and Vaina 1982). Conceptual structures of information about natural actions include function-argument structures which are ultimately linked to spatial structure encoding. The spatial structure encoding linked to conceptual structures may serve as a representation of what is distributed in gesture as complementary information. That is, quasi-geometric information about natural actions is conveyed as complementary meaning in the gesture modality.

The utterance “the car spun” illustrates the natural actions hypothesis. There is no implicit path in the meaning of this sentence. What the sentence describes is the internal motion of the subject without any implication with respect to the subject’s location or with respect to any other object. The verb “spun” expresses an idiosyncratic manner of motion that is not decomposable into algebraic features in a way that distinguishes spinning from other possible motions. Its conceptual representation is of the function-argument structure of type (6) which uses the generic MOVE to encode verbs of manner of motion. After argument substitution the conceptual structure takes the form of (7). If a gesture accompanies this sentence, it is likely to include a hand rotating around a vertical axis which corresponds to the natural action of spinning.

(6) (MOVE ((THING)))

(7) (MOVE (car))

4.3. The Marked Features Hypothesis

Cassell and Prevost suggest that redundancy across speech and gesture marks less predictable information (Cassell & Prevost 1996, Cassell & Prevost under review). By appealing to this notion, I claim that a marked feature in a conceptual structure serves as a representation of what is distributed in gesture as redundant information. In Jackendoff’s conceptual structures, the treatment of verbs and prepositions entails developing the primitive semantic fields into coherent feature systems. For example, a coherent feature system for location and contact establishes distinctions between marked and unmarked cases of location and contact. The marked case of a feature is the particular case of the feature. The unmarked case of a feature is the ordinary case of the feature. Marked features may serve as a representation of what is conveyed in gesture as redundant information. That is, marked cases of the feature system are distributed as redundant meaning in the gesture modality.

The utterance “the trailer was attached to the cab” illustrates the marked features hypothesis. The location of the trailer is not the ordinary case of being at or next to a reference entity. In this case, the trailer is located in contact with the reference entity. And this is not an ordinary type of contact but one in which the trailer is joined to the cab. Its conceptual representation, which marks location and contact, is of the function-argument structure of type (8) where “a” is a subscript that stands for attachment as a marked case of contact. After argument substitution the conceptual structure takes the form of (9) . If a gesture accompanies this sentence, it is likely to include a spatial arrangement of two hands in which at one moment the two hands are in contact, corresponding to a demonstration in the gesture space of a marked case of location and contact.

(8) (BEa ((THING)
(ATa (THING))))

(9) (BEa ((trailer)
(ATa (cab))))

4.4. Analysis of Experimental Data and Representations of Specific Distributions

Table 1: Specific distribution representations vs. presence or absence of gesture

gesture	implicit arguments	natural actions	marked features
presence	10/12 (84%)	4/5 (80%)	10/13 (77%)
absence	2/12 (16%)	1/5 (20%)	3/13 (23%)

Table 1 shows the distribution of gestures with respect to the hypotheses in the sample of experimental data. The presence percentages encompass those instances in which, given a representation of a specific distribution, what is indicative in the representation of demanding expression in gesture turns out to be correspondingly conveyed in gesture. The absence percentages encompass those instances in which, given a representation of a specific distribution, what is indicative in the representation of demanding expression in gesture turns out not to be correspondingly conveyed in gesture. The instances of implicit arguments all occurred in external motion descriptions. The instances of natural actions all occurred in internal motion descriptions. The instances of marked features all occurred in descriptions of contact and location. The following sections discuss instances in the sample of experimental data that support or do not support the three hypotheses about the distribution of information across speech and gesture.

4.5. Implicit Arguments: External Motion Information

The implicit arguments hypothesis is that an implicit argument in a conceptual structure may serve as a representation of what is distributed in gesture as complementary information. Cases of implicit arguments in the analyzed sample of experimental data were instances in which the information missing from speech was implicit information about external motion. The distinction of external motion serves to point to the motion of an entity as a whole, its traversal in a region (Jackendoff 1987). Using a reference object, a region is defined to locate the entity in motion. Prepositions, as linguistic elements that express spatial relations, map the reference object to a path. The geometry of a path is sometimes specified by prepositions denoting its orientation, such as “along,” “across,” or “around,” or its distribution in a region, such as “all over” and “throughout” (Jackendoff & Landau 1991). The implicit arguments hypothesis is supported by 84% of the implicit argument instances of the analyzed sample of experimental data. Two examples of these instances are “the coyote comes out” and “the road runner returns.” For both of these the missing information is available from the context. The information exhibited by gesture may be an aid for conversants to reduce the processing time to retrieve information that they know or could infer (Walker 1993).

In the case of the utterance “the coyote comes out”, its conceptual representation is of the function-argument structure of type (4) which after argument substitutions takes the form of (10). The gesture accompanying the sentence was a left hand upwardly moving from left waist to center upper chest. There is external motion information missing from speech because “comes out” implies the traversal of a path from inside an unknown place. The path is unspecified because the speech of the speaker is not locating the traversal of the moving entity, the coyote, with respect to other object. The narrator provides complementary information in gesture by exhibiting a vertical axis of movement elongation that conveys an upward orientation in the traversal of an unspecified path.

(10) (GO ((coyote)
(FROM (INSIDE (void)))))

In the case of the utterance “the road runner returns” its conceptual representation is also of the function-argument structure of type (4) which after argument substitution takes the form of (11). The gesture accompanying this sentence was a right hand downwardly moving from right upper shoulder to center lower chest. There is external motion information missing from speech because “returns” implies the traversal of a path. The path is unspecified because the speech of the speaker is not explicitly locating the traversal of the moving entity, the road runner, with respect to the other object. The narrator provides complementary information in the gesture by exhibiting a vertical axis of movement elongation that conveys a downward orientation in the traversal of an unspecified path.

(11) (GO ((road-runner)
(TO (void))))

In the other 16% of the instances, utterances containing implicit information about external motion do not support the implicit arguments hypothesis. An example of this is an utterance about a rock that “starts rolling.” There is external motion information missing because rolling implies the traversal of a path. The gesture accompanying this utterance was a left hand in a circular motion from near-the-body at the center of the upper chest to away-from-the-body parallel to the right shoulder. This gesture resembles the natural action of rolling more than any expression of complementary information about the traversal of an unspecified path. One could argue that the speech-gesture production system favors the surface manifestation of what is most salient in the immediate context of the scene. In this case the natural action of rolling could have been favored over the expression of external motion information because the speaker intended to depict the beginning of a type of internal motion and not the traversal of a path. The next section discusses also this type of dual presence when it attempts to explain why some instances do not support the natural actions hypothesis.

4.6. Natural Actions: Internal Motion Information

The natural actions hypothesis is that a spatial structure encoding linked to a conceptual structure of a natural action may serve as a representation of what is distributed in gesture as complementary information. Cases of natural actions in the analyzed sample of experimental data were instances of internal motion information. The distinction of internal motion serves to point to those motions of an entity that are object-centered, internal dispositions such as bouncing, waving or spinning (Jackendoff 1987, Jackendoff 1990). The natural actions hypothesis is supported by 80% of the instances of the analyzed sample of experimental data. One example of these instances is the utterance “he unravels,” where unravels is an intransitive verb.

The conceptual representation for the utterance “he unravels”, is of the function-argument structure of type (6). After argument substitution the conceptual structure takes the form of (12) with a link to a spatial encoding of the natural action of unraveling. The gesture accompanying this sentence was a spatial arrangement of both hands, parallel to each other, with extended index fingers, pointing at each other, and a small space between them, rotating in a counterclockwise direction around an horizontal axis in front of center lower chest. The narrator provides complementary information in the gesture by exhibiting the way the coyotes’s elastic body disentangles through several cycles of self-unwrapping.

(12) (MOVE (coyote))

In the other 20% of the instances, utterances about internal motions do not support the natural actions hypothesis. An example of this is the utterance “the rock is rolling.” The internal motion of rolling is a natural action that may imply the traversal of a path if it takes place over a surface. Other internal motions that are natural actions always seem to imply the traversal of a path such as walking or running. For these the gesture seems to convey the information about external motion. An example of this was an utterance that described “the road runner is running.” The gesture that accompanied that sentence was a rectilinear right hand movement from right shoulder to center upper chest. In the case of “the rock is rolling” the gesture that accompanied the sentence was a rectilinear left hand movement from left shoulder to center upper chest. Those gestures do not resemble the internal motions, neither running or rolling. They do resemble the traversal of a path in a particular direction over a particular axis of movement elongation. If this dual presence of external and internal motion information occurs, the speech-gesture production system seems to favor the surface manifestation of external motion information unless the speaker does not intend to depict the whole motion.

4.7. Marked Features: Contact and Location

The marked features hypotheses is that a marked feature in a conceptual structure may serve as a representation of what is distributed in gesture as redundant information. Cases of marked features in the analyzed sample of experimental data were instances of marked location or contact. The marked features hypothesis is supported by 77% of the marked feature instances of the analyzed sample of experimental data. An example of these instances is the utterance “he hits the ground,” which is a marked case of location.

The conceptual representation for the utterance “he hits the ground” is of the function-argument structure of type (13) where “c” is a subscript that stands for contact as a marked case of location. After argument substitution the conceptual structure takes the form of (14). The gesture accompanying the sentence was a right hand at a fixed position in front of the center lower chest and a left hand moving from left shoulder to center lower chest until it impacts with the right hand. This is a marked case of location because the thing in motion not only ends up at a place with reference to another entity but it is also in contact with the reference entity. The narrator provides redundant information in gesture by exhibiting two entities coming in contact.

(13) (BEc ((THING)
              (ATc (THING))))

(14) (BEc ((coyote)
              (ATc (ground))))

In the other 23% of the instances, utterances with marked cases of contact and location do not support the marked features hypothesis. An example of this is an utterance describing that a rock "hits another side." The gesture accompanying this utterance was a right hand in a curved movement from center upper chest to right shoulder. This gesture does not resemble contact as a marked case of location. The gesture conveys information about the traversal of an unspecified path by a rock that is in motion and ends up in contact with an unknown reference object. It appears that hitting is not what the speaker is intending to depict. The use of "another" could indicate that the speaker intends to contrast that the location that the rock contacts in this impact is different from the location of the previous impact.

5. Realization of Semantic Information in Gestural Forms

In this chapter, I offer three hypotheses, as presented in (1), (2), and (3), regarding the realization of semantic information in gestural forms.

(1) The shape hypothesis is that distinctions of meaning about the shape of an object are realized in gesture as hand shapes.

(2) The trajectory hypothesis is that distinctions of meaning about the trajectory of an object are realized in gesture as shapes of hand motion.

(3) The interaction dyad hypothesis is that semantic distinctions about the interaction dyad of force dynamics are realized in the arrangement of the number of hands.

5.1. The Shape Hypothesis

Kendon suggests that what is difficult to express in speech may be conveyed by gesture, including spatial information that is elusive to speech (Kendon 1994, Cassell et. al. 1994b). By appealing to this notion, I claim that distinctions of meaning about the shape of an object are realized in gesture as the shape of the hand. Shape is an important element in the criteria for identification of object categories (Landau et. al. 1988). Many objects are linguistically categorized on the basis of appearance. The translation of object shape descriptions into language is very difficult because language cannot convey all the richness of spatial representations of complex contours and edge patterns (Jackendoff & Landau 1991).   The relative sizes and proportions that characterize a shape are geometric notions far more suitably spelled out in a spatial representation than in language. Several researchers have proposed visual representations of the geometric properties of physical objects (Marr 1982, Biederman 1987). In these representations objects are decomposed into parts and spatial axes that organize the parts of the objects and relations among them (Jackendoff & Landau 1991). The hand shape of the gesture may express semantic distinctions between objects of round shape and objects of non-round shape.

The hand shape of a gesture that accompanies the utterance “a ball struck the glass” illustrates the shape hypothesis. If a gesture accompanies this sentence, it is likely to include two hands, a hand moving until it makes contact with the other hand at a fixed position. The shape of the hand in motion is likely to have round features in contrast with the other hand that represents a non-round object using other idiosyncratic features.

5.2. The Trajectory Hypothesis

Jackendoff and Landau suggest that language is crude in expressing information about distance and orientation, unless the speaker adopts a culturally stipulated system of measurements, which implies that language cannot convey many details of spatial representations (Jackendoff & Landau 1991). By appealing to this notion, I claim that distinctions of meaning about the trajectory of an object are realized in gesture as the shape of the hand motion. Jackendoff’s conceptual structures use the PATH conceptual category to specify the trajectory of a thing in motion (Jackendoff 1990, Jackendoff & Landau 1991). Language specifies these trajectories using prepositions and verbs that express spatial relations. For example, “to run by” involves traversing a trajectory that at some point is near to some place. Another example is “to go through” which involves traversing a trajectory that at some point is at some place. Other prepositions, such as “along,” “across,” and “around,” construct a path for a thing in motion that traverses a linear region in a trajectory that is coaxial with the region. The PATH conceptual structure allows the specification of a trajectory that terminates or begins at a region using operators TO and FROM. Prepositions, such as “to,” “into,” and “onto,” express a trajectory that terminates at, in, and on some place respectively. The preposition “from,” a reversal of the previous case, expresses a trajectory that begins at a place. The shape of the hand motion may express semantic distinctions between paths with endpoints and paths without endpoints.

The shape of the hand motion accompanying the utterance “the car came out of the parking lot” illustrates the trajectory hypothesis. If a gesture accompanies this sentence, it is likely to include a gesture with a curved hand trajectory.

5.3. The Interaction Dyad Hypothesis

Kendon suggests that what is difficult to be expressed in speech may be conveyed by gesture, including dynamic features elusive to speech (Kendon 1994, Cassell et. al. 1994b). By appealing to this notion, I claim that semantic distinctions about interaction dyads of force dynamics are realized in the arrangement of the number of hands. Force dynamics involve the interaction of two protagonists. Jackendoff's conceptual structures use the AFF function, which stands for affect, to encode the dyadic relations of the protagonists according to Talmy's theory of force dynamics configurations (Jackendoff 1990, Talmy 1988). If the dyadic relation is one of opposition, one of the protagonists is the agonist and the other is the antagonist. The agonist has a tendency to perform or to not perform an action. The antagonist opposes the tendency. If the dyadic relation is one of cooperation both protagonists strive for the same potential effect. The arrangement of the number of hands may express semantic distinctions between the dyadic relations of opposition and cooperation.

The arrangement of the number of hands accompanying the utterance "the truck hits the car" illustrates the interaction dyad hypothesis. If a gesture accompanies this sentence, it is likely to include an arrangement of two hands in which one hand, the truck as the antagonist, comes in contact with the other hand, the car as the agonist.

5.4. Analysis of Experimental Data and Realization of Semantic Distinctions

Table 2: Semantic distinctions realizations vs. absence or presence of gestural forms

gestural form	shape	trajectory	interaction dyad
presence	8/11 (73%)	6/8 (75%)	7/10 (70%)
absence	3/11 (27%)	2/8 (25%)	3/10 (30%)

Table 1 shows three semantic distinction realizations analyzed in the sample of experimental data. The presence percentages encompass those instances in which, given a type of semantic distinction, this distinction is correspondingly realized in a particular gestural form. The absence percentages encompass those instances in which, given a type of semantic distinction, this distinction is not correspondingly realized in a particular gestural form. The instances of shape were semantic distinctions of roundness as an easily isolated shape feature. The instances of trajectory were semantic distinctions of traversals with an endpoint. The instances of interaction dyads were semantic distinctions of dyadic relations of opposition as the only dyadic relation present in the sample of experimental data. The following sections discuss instances in the sample of experimental data that support or do not support the three hypotheses about the realization of semantic information in gesture.

5.5. Shape: Roundness

The shape hypothesis is that distinctions of meaning about the shape of an object are realized in gesture as hand shapes. Cases of shape in the analyzed sample of experimental data were instances in which the shape of the object was round. The shape hypothesis is supported by 73% of the instances. An example of these instances is the utterance “the rock is rolling.” The hand shape of the gesture accompanying this sentence was similar to the ASL hand shape for O, which could be described as a hand with curled fingers. The narrator exhibits the shape, of the thing in motion, in the gesture by using a hand shape with round features.

In the other 27% of the instances, utterances containing an entity with a round shape do not support the shape hypothesis. An example of this is the utterance “a rock falls.” The gesture accompanying this utterance is a spatial arrangement of two hands coming in contact with each other. The right hand at a fixed position, center upper chest, and the left hand moving from left shoulder, representing the rock, with a hand shape similar to the ASL hand shape for 5, which could be described as a hand with a flat palm. The flat hand shape does not correspond to the round shape of the rock. It appears that the speaker intends to depict an aspect of the impact of the rock with the surface or the opposition in force dynamics and therefore there is a blurring of the shape of the rock with other semantic information.

Almost all of the hand shapes in the analyzed sample of experimental data can be divided in three groups. First, closed palms, ranging from curled fingers to fists, resembling round features (These ones were similar to ASL hand shapes for C, bent 5, A, S, O and baby O). Second, flat palms, resembling level or smooth features (These ones were similar to ASL hand shapes for B, B spread and 5). Third, palms with extended index fingers, resembling thin features (These ones were similar to ASL hand shapes for G,L and H).

5.6. Trajectories: Traversals with Endpoints

The trajectory hypothesis is that distinctions of meaning about the traversal of an object are realized in gesture as shapes of hand motion. Cases of trajectory in the analyzed sample of experimental data were instances in which the traversals had an endpoint. Endpoint is the term to point out to the extremes, beginning or end, of an interval of the action of traversing a region. The trajectory hypothesis is supported by 75% of the instances. An example of these instances is the utterance “the catapult goes out of the ground.” The gesture accompanying this sentence had a curved hand trajectory. The narrator exhibits in the gesture the traversal of the catapult through a trajectory with an endpoint, from the ground, by using a curved hand trajectory movement.

In the other 25% of the instances, utterances containing a trajectory with an endpoint do not support the trajectory hypothesis. An example of this is the utterance about the road runner that “goes into the painting.” The hand trajectory of the gesture accompanying this utterance was rectilinear. The rectilinear hand trajectory does not correspond to a trajectory with an endpoint, the painting. An attempt to explain this resides in considering that the preposition “into” has several readings. In one of its readings, it not only means to go to a reference object, the painting which is an endpoint in the path, but also to traverse the interior of that reference object. It could be argued that the speaker intended to depict only the traversal of the interior of the painting instead of also depicting the traversal to the painting. The preposition “into” has another reading which is related to the action of coming in contact with the reference object under a significant amount of force. This reading was, of course, the applicable one when the narrator described what happened to the coyote by going into the painting.

5.7. Interaction Dyad: Relations of Opposition

The interaction dyad hypothesis is that distinctions about the interaction dyad of force dynamics are realized in the arrangement of the number of hands. Cases of interaction dyads in the analyzed sample of experimental data were instances in which the dyadic relation was one of opposition. The interaction dyad hypothesis is supported by 70% of the instances. An example of these instances is an utterance describing that the coyote "pulls the pin out." The gesture accompanying this sentence included an arrangement of two hands in which the left hand, coyote's hand as the antagonist, comes in contact with the right hand, the hand grenade with a pin as agonist. The narrator exhibits in the gesture the force dynamics of resistance with contact, or close proximity, depicting opposition.

In the other 30% of the instances, utterances containing opposition in force dynamics do not support the interaction dyad hypothesis. An example of this is the utterance "he loosens the anvil." The one hand movement in the gesture that accompanied this sentence did not exhibit the dyadic relation of opposition. The right hand moved away from the body to possibly convey information about the outward trajectory of the anvil after untying it. An attempt to explain this resides in considering that there could be several readings of the utterance. In one of its reading "loosing the anvil" could mean to unload the anvil instead of to untie the anvil. One could argue that the speaker intended to depict the unburdening of the anvil from the air balloon instead of depicting the force dynamics of untying the anvil.

6. Semantically Appropriate Gestures for Embodied Language Generation

This chapter presents an overview of the prototype. The first section describes the prototype through a description of the layout of its components. The second section discusses the scope of the implementation for each one of its components. The third section summarizes its limitations.

6.1. Prototype Layout Description

Figure 1 presents a layout of the prototype components. The inputs to the utterance planner are conceptual structures as abstract semantic representations. Appendix A contains a sample of these conceptual structures as prototype inputs. These abstract semantic representations capture state and event attributes relevant to descriptions of motion scenes. An appropriate encoding scheme for these abstract semantic representations was adopted from Jackendoff's conceptual semantics because of its provisions for primitives that capture generalizations about action descriptions and reveal what semantic information is missing from, and what semantic information is marked in, the content of speech (Jackendoff 1987, Jackendoff 1990).

The utterance planner extracts semantic information from conceptual structures and distributes their mapping into highly constrained sentential and gestural specifications. The mapping into sentential specifications is a conversion from identified types of sentences to an arrangement of specifications in the FUF (Functional Unification Formalism) attribute-value language format (Elhadad 1993). Appendix C contains a sample of these lexicalized specifications as inputs to the FUF surface generator. FUF is a natural language generator program that uses the technique of unification grammars. Its two main components are a unifier and a linearizer. The input to the unifier is a semantic description of the text to be generated and a unification grammar. The output of the unifier is a syntactic description of the text. The syntactic description is the input to the linearizer which produces as output an English sentence.

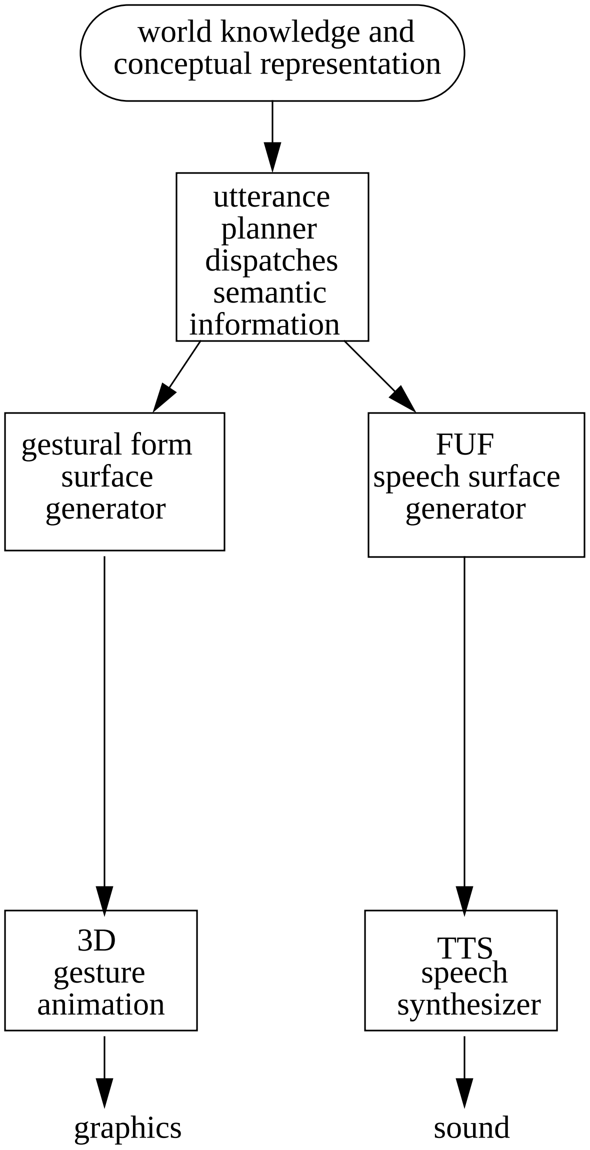

world knowledge and conceptual representation
utterance planner dispatches semantic information
gestural form surface generator
FUF speech surface generator
3D gesture animation
TTS speech synthesizer
graphics
sound

Figure 1. Layout of components of prototype to realize descriptive animated utterances

The mapping into gestural specifications is a transformation from representations of semantic information distribution and distinctions to sequences of gesture forms in specifications of motor articulations for hand shapes and trajectories. The utterance planner sends its output of lexicalized or gestural representations into the corresponding surface generator. Appendix B contains a sample of these gestural representations as inputs to the prototype animation program.

The inputs to the FUF speech surface generator are general propositions in an attribute-value language format that are transformed into strings of words using SURGE, a syntactic realization component for natural language generation systems (Elhadad 1992, Elhadad & Robin 1996). SURGE is a comprehensive grammar of the English language. The inputs to the gesture surface generator are general hand movement attributes as pairs of sequences of motor articulations that are transformed into graphical animation commands. These commands specify the hand shape and trajectory of particular gestures. Appendix D contains frontal and side views of the upper-body figure of the prototype graphical animation program. The speech synthesizer uses Entropic's TTS (TrueTalk System) program. TTS is a text-to-speech application program.

The domain for motion events is car accident descriptions. This domain was chosen because of its simple and repetitive structure of states and events. It is also spatially concrete. It has potential to evoke iconic gestures since it is rich in images abundant in dynamic features, elements, and properties typical of a motion scene. Research shows that there is more display of gesture when the verbal task entails mental images of action and objects that include visual, spatial, and motoric information (Rime & Schiaraturea 1991). The inputs to this prototype are conceptual representations of a car accident state or event description. As an output it produces an animated descriptive utterance through the coordinated animation of gesture realization in an upper-body figure and synthesis of speech.

6.2. Main Prototype Components

The utterance planner is mainly a dispatcher of semantic information to the gesture surface generator. Its inputs are conceptual representations. It can parse Jackendoff's most important types of function-argument structures. The utterance planner associates the conceptual representation with a type of sentence. After identifying the type of sentence to be generated, the utterance planner fills out the slots of an intermediate sentential representation, which is a lexicalized specification in the FUF attribute-value language format, for the realization of the sentence. This is done using the information obtained in the parsing of the conceptual structure. The utterance planner analyses the conceptual structure specification to associate it with sequences of gesture forms, hand shapes and trajectories, corresponding to the distribution and realization of semantic information. The sequences of gestural forms to be distributed exhibit redundant information, marking location and contact, and complementary information, specifying information about the trajectory of a thing in motion. The sequences of gestural forms to be realized exhibit information about the shape of an entity (marking roundness) and force dynamics (marking opposition). After identifying the types of sequences of gestural forms, the utterance planner fills out the slots of an intermediate gestural specification which is used in the gesture surface generator for the realization of the gesture. The utterance planner uses the information obtained from the parsing and from the semantic analyses of the conceptual structure or from the retrieval of domain-dependent information from the knowledge base.

The gesture surface generator maps the intermediate gestural specification of sequences of gestural forms into graphical animation specifications for the motor articulation of hand shapes and trajectories. For every movement sequence there is a specification for an initial and final position. The graphical command specifications define handness (either left or right), hand shape (ASL shapes for either A,S, O, baby O, B, B spread, C, 5, bent 5, G, L, or H), palm normal (either up, down, away from body, toward, left, or right), palm approach (same values as palm normal), position in a plane of gesture space (either left shoulder, right shoulder, upper center chest, left belly, center belly, or right belly), plane of gesture space (either away from the body, center, or toward the body), and duration of movement (seconds as time measurement).

6.3. Overall Prototype Limitations

The prototype only realizes certain types of animated utterances. Its extensibility for including new types of sentences is dependent on matching a conceptual structure with an intermediate lexicalized representation in FUF format. The prototype graphical animation program allows only static hand shapes and rectilinear movements from one fixed point to another fixed point in the gesture space. Its extensibility for including dynamic hand shapes, curved hand trajectories, and graphical animation of natural actions is dependent on the implementation of a model of inverse kinematics for motor articulation. The graphical program only animates certain hand shapes but it allows the definition of new hand shapes using a set of features to decompose the hand configuration in parameters, such as palm (flat, open, or closed), fingers (extended or non-extended), thumb position (in, out, or over) and degree of finger spread (together, relaxed, apart).

7. Conclusion

This thesis has presented a combined approach of theoretical and practical methodologies to study the semantic relation between speech and gesture. Previous work does not specify what semantic information is realized in gesture and what is realized in speech, nor how this distribution may be represented. This thesis has introduced a set of hypotheses to explain the distribution and realization of semantic information across speech and gesture. It has extended an existing body of work on the rule-based generation of semantically appropriate gestures in embodied language generation by proposing hypotheses and building a prototype as a vehicle to experiment with the applicability of some of the proposed hypotheses and to explore their usefulness and suitability for producing semantically appropriate gestures in embodied language generation.

The set of six hypotheses concerning the distribution and realization of semantic information across speech and gesture are:

(1) The implicit arguments hypothesis is that an implicit argument in a conceptual structure may serve as a representation of what is distributed in gesture as complementary information.

(2) The natural actions hypothesis is that a spatial structure encoding linked to a conceptual structure of a natural action may serve as a representation of what is distributed in gesture as complementary information.

(3) The marked features hypothesis is that a marked feature in a conceptual structure may serve as a representation of what is distributed in gesture as redundant information.

(4) The shape hypothesis is that distinctions of meaning about the shape of an object are realized in gesture as hand shapes.

(5) The trajectory hypothesis is that distinctions of meaning about the trajectory of an object are realized in gesture as shapes of hand motion.

(6) The interaction dyad hypothesis is that semantic distinctions about the interaction dyad of force dynamics are realized in the arrangement of the number of hands.

Suggestions of extensions to this work reside in considering its limitations and range of applicability. The analyzed sample of experimental data supports the proposed hypotheses. However, the selected sample of experimental data only includes instances in which iconic gestures were present. It is not representative of absence of gesture. Future efforts to refine the hypotheses may involve selecting a random sample corpus that has instances of the absence of gesture.

The built prototype embodies some of the proposed hypotheses. This makes possible to test and evaluate them with controlled experiments. The built prototype may be expanded to embody the natural actions and trajectory hypotheses. A general approach to design a controlled experiment using the built prototype may include a procedure in which subjects are presented with randomly generated gestures and gestures generated according to the proposed hypotheses. Afterwards, the subject may answer a questionnaire to be used for assessing the usefulness and suitably of the proposed hypotheses. This may be the beginning of future research efforts for evaluating the proposed hypotheses.

Future work includes investigating the use of conceptual semantics and the results of this thesis outside the domain of motion events. Discussions about the link between gesture and action and the claim that the representation of semantic features in iconic gestures is acquired through action schemata present a point of departure for understanding more the applicability of the results of this thesis beyond motion events to include other semantic fields (Cassell in press).

Future work also includes addressing other aspects of the work that remain unexplored such as examining points of contact between conceptual structures and information structure heuristics to predict when complementary and redundant information across speech and gesture occurs. This work could be applied to embodied language beyond utterance generation to be included in monologue and dialogue generation. It could also be applied to gesture recognition as part of investigating the role of a conceptual semantics in gesture interpretation.

Appendix A: Sample of Conceptual Representations as Inputs to Prototype

Pairs of utterances followed by the corresponding conceptual structures.

“The truck comes out”

```
(1)  ( GO ( ( truck )
                ( FROM ( INSIDE ( void ) ) ) ) )
```

“The trailer is attached to the cab”

```
(2)  ( BEa ( ( trailer )
                 ( ATa ( cab ) ) ) )
```

“The truck left”

```
(3)  ( GO ( ( truck )
                ( FROM ( void ) ) ) )
```

“The car hit the tree”

```
(4)  ( BEc ( ( car )
                 ( ATc ( tree ) ) ) )
       ( AFF- ( ( car ) ( tree ) )
```

“The car goes through”

```
(5)  ( GO ( ( car )
                ( VIA ( INSIDE ( void ) ) ) ) )
```

“The truck hits the car”

```
(6)  ( BEc ( ( truck )
                 ( ATc ( car ) ) ) )
      ( AFF- ( truck ) ( car ) )
```

“The car goes by”

```
(7)  ( GO ( ( car )
                ( VIA ( NEAR ( void ) ) ) ) )
```

“The ball hits the hood”

```
(8)  ( BEc ( ( ball )
                 ( ATc ( hood ) ) ) )
       ( AFF- ( ( ball ) ( hood ) ) )
```

“The car enters”

```
(9)  ( GO ( ( car )
                ( TO ( INSIDE ( void ) ) ) ) )
```

Appendix B: Sample of Gesture Specifications as Inputs to the Gesture Animation

Figures 1 and 2 show gesture animation frames with the corresponding specifications (1) and (2).

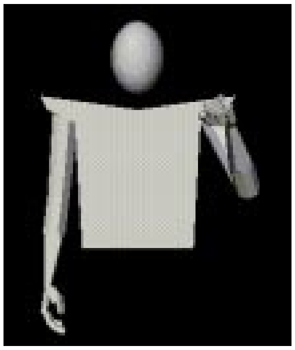


Figure 1. Preparation phase for “the ball hits the hood”

(1)

l

hS

comfort

comfort

upperLeftShoulder

center

linear

1

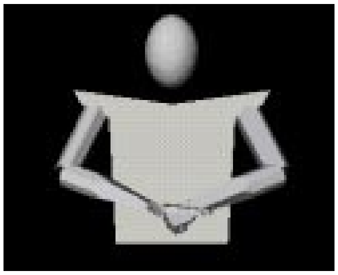


Figure 2. Stroke phase for “the ball hist the hood”

(2)
r

h5

up

comfort

belly

center

linear

1

l

hS

comfort

comfort

belly

center

linear

1

Appendix C: Sample of Lexicalized Specifications as Inputs to FUF

```
(def-test j17
  “The truck pushes the car.”
  ((cat clause)
   (proc ((type material)
          (lex “push”)))
   (partic ((agent ((cat common) (lex “truck”)))
            (affected ((cat common)
                       (lex “car”)))))))

(def-test j1.1
  “The truck comes out.”
  ((cat clause)
   (tense present)
   (proc ((type composite)
          (agentive no)         (relation-type locative)
          (lex “come”)))
   (partic ((affected ((cat common) (lex “truck”)))
            (located {^ affected})
            (location ((cat adv) (lex “out”)))))))

(def-test j2
  “The truck left.”
  ((cat clause)
   (tense past)
   (proc ((type material) (agentive no) (lex “leave”)))
   (partic ((affected ((cat common) (lex “truck”)))))))

(def-test j18
  “A trailer is attached to the cab.”
  ((cat clause)
   (process ((type locative) (mode equative)
             (lex “attach”) (voice passive)
             (passive-prep ((lex “to”)))))
   ;; Default is without agent - override it.
   (agentless no)
   (partic ((location ((cat common)
                       (definite no)
                       (lex “trailer”)))
            (located ((cat common)
                      (lex “cab”)))))))
```

Appendix D: Sample of Transcriptions for the Analysis of Experimental Data

20:56:31.09 “[pulls] the pin out” 20:56:32.12

hands:
R, close to C, away, left
L, close to O, away, right

motion:
R, right waist, fixed
L, from right waist to left waist, curved

meaning:
R, hand grenade, CPV, spherical hand shape
L, fact of pulling, CPV, curved motion of pulling

20:32:02.08 “he [unravels]” 20:32:04.10

hands:
L, close to L, right, toward
R, close to L, left, toward

motion:
L and R parallel to each other with a small space between
them rotating around an horizontal axis in front of center waist

meaning:
L and R, cyclical manner of motion, OPV, counterclockwise rotation

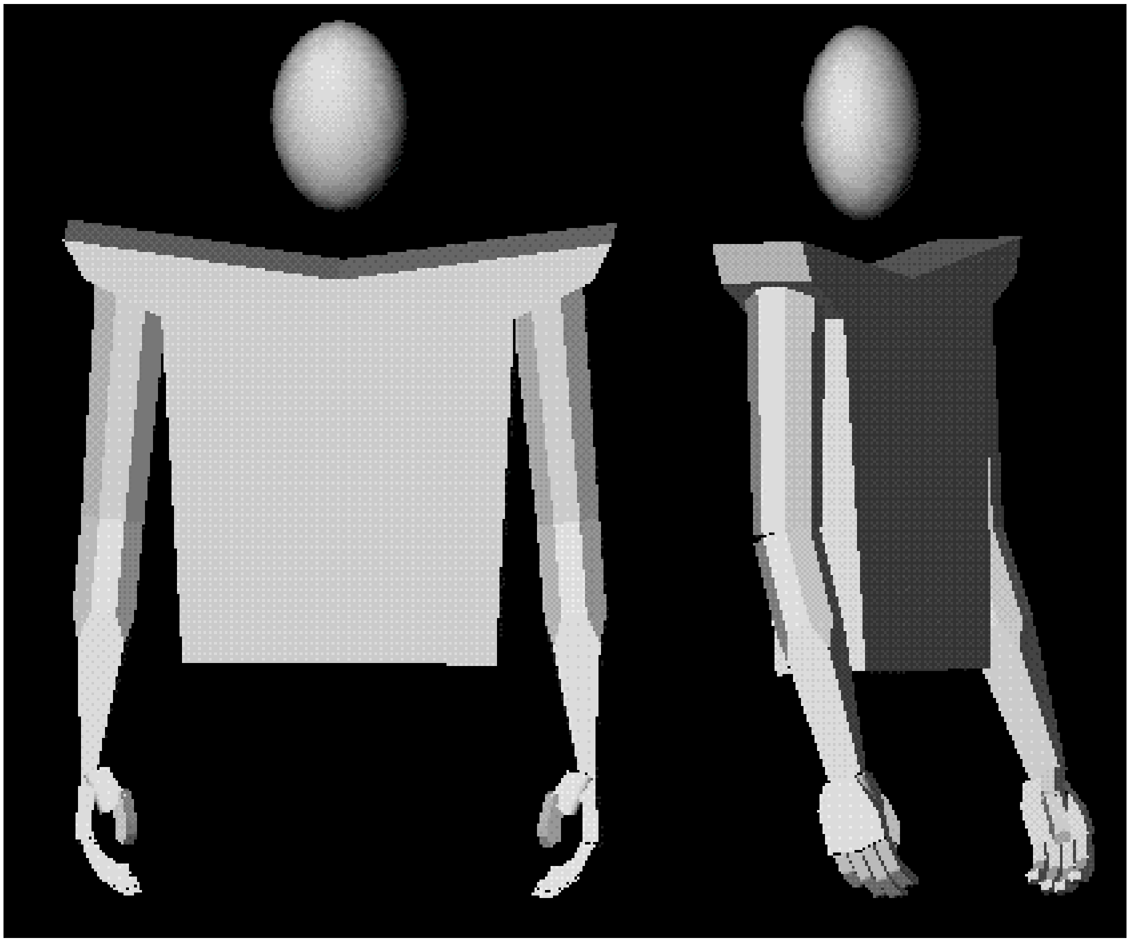
21:01:16.28 “the road runner [returns]” 21:01:18:03

hands:
R, close to 5, up, toward

motion:
R, from right shoulder to center waist, curved

meaning:
R, downward traversal of a path, OPV, vertical axis of movement elongation

Appendix E: Frontal and Side Views of Gesture Animation Upper-Body Figure



References

Bers, J. (1995). Directing animated creatures through gesture and speech. M.S. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Biederman, I. (1987). Recognition by components: a theory of human image understanding. Psychological Review 94, 115-147.

Bolt, R.A. (1980). “Put-That-There”: voice and gestures at the graphics interface. Computer Graphics, 14(3), 262-70.

Bolt, R.A. (1987). The integrated multi-modal interface. Transactions of the Institute of Electronics, Information and Communication Engineers (Japan), J79-D(11), 2017-2025.

Bolt, R.A. & Herranz, E. (1992). Two-handed gestures in multi-modal natural dialog. Proceedings of OISI ‘92, Fifth Annual Symposium on User Interface Software and Technology, Monterey, CA.

Cassell, J & McNeill, D. (1991). Gesture and the poetics of prose. Poetics Today, 12(3): 375-404.

Cassell, J., Pelachaud, C., Badler, N.I., Steedman, M., Achorn, B., Beckett, T., Douville, B., Prevost, S. & Stone, M. (1994a). Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. Computer Graphics 94.

Cassell, J., Steedman, M., Badler, N., Pelachaud, C., Stone, M., Douville, B., Prevost, S., &Achorn, B. (1994b). Modeling the interactions between gesture and speech. In Proceedings of the Cognitive Science Society Annual Conference, Atlanta, GA.

Cassell, J. & Prevost S. (1996). Distribution of semantic features across speech and gesture by humans and computers. Proceedings of the Workshop on the Integration of Gesture in Language and Speech. Wilmington, DE.

Cassell, J. McNeill, D. & McCuloguh, K.E. (in press). Speech-gesture mismatches: evidence for one underlying representation of linguistic and non-linguistic information. Cognition.

Cassell, J. (in press). A framework for gesture generation and interpretation. In R. Cipolla and A. Pentland (eds.), Computer Vision in Human-Machine Interaction. Cambridge University Press.

Cassell, J. & Prevost S. (under review). Embodied natural language generation: a framework for generating speech and gesture.

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

Cohen, A.A. (1977). The communicative function of hand illustrators. Journal of Communication, 27(4): 45-63.

Cohen, A.A. & Harrison, R.P. (1973). Intentionality in the use of hand illustrators in face-to-face communication situations. Journal of Personality and Social Psychology, 28, 276-279.

Ekman, P. and W. Friesen (1969). The repertoire of nonverbal behavioral categories-origins, usages and coding. Semiotica, 1:49-48.

Elhadad, M. (1993). FUF: the universal unifier - user manual, version 5.2. Technical Report CUCS-038-91, Columbia University, New York.

Elhadad, M. & Robin J (1996). An overview of SURGE: a reusable comprehensive syntactic realization component. Technical Report 96-03, Mathematics and Computer Science Department, Ben Gurion University in the Negev, Israel.

Elhadad, M., McKeown K. & Robin J. (1997). Floating constraints in lexical choice. Computational Linguistics 23 (2), 195-239.

Fillmore, C. J. (1982). Frame semantics. In Linguistic Society of Korea (eds.), Linguistics in the Morning Calm. Hanshin Publishing Co.

Goldin-Meadow S., Alibali M., Church B. (1993). Transitions in concept acquisition: using the hand to read the mind. Psychological Review, 100 (2), 279-297.

Gruber, J.S. (1965). Studies in lexical relations. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge.

Halliday, M. (1967). Intonation and grammar in British English. Mouton.

Jackendoff, R. (1987). On beyond zebra: The relation of linguistic and visual information. Cognition, 26, 89-114.

Jackendoff, R. (1990). Semantic structures. MIT Press.

Jackendoff, R. & Landau, B. (1991). Spatial language and spatial cognition. In D. Napoli & J. Kegl (eds.), Bridges between Psychology and Linguistics. Hillsdale.

Katz, J. J. & Fodor, J. A. (1963). The structure of a semantic theory. Language, 39: 170-210.

Kendon, A. (1972). Some relationships between body motion and speech. In A.W. Siegman & B. Pope (eds.), Studies in Dyadic Communication. Pergamon Press.

Kendon, A. (1980). Gesticulation and speech: two aspect of the process of utterance. In M.R. Key (ed.), The Relation Between Verbal and Nonverbal Communication. Mouton.

Kendon, A. (1986). Current issues in the study of gesture. In J. Nespoulous, P. Perron & A. Lecours (eds.), The Biological Foundations of Gestures: Motor and Semiotic Aspects. Hillsdale.

Kendon, A. (1994). Do gestures communicate: A review. Research on Language and Social Interaction, 27.

Koons, D., Sparrell, C. & Thorisson, K (1993). Integrating simultaneous input from speech, gaze and hand gestures. In M. Maybury (ed.), Intelligent MultiMedia Interfaces. Cambridge, MA.

Koons D. (1994). Capturing and interpreting multi-modal descriptions with multiple representations. AAAI Symposium on Intelligent Multi-Modal Multi-Media Interface Systems. Stanford, CA.

Landau, B., Smith, L., & Jones, S. (1988). The importance of shape in early lexical learning. Cognitive Development, 3, 299-321.

Lasher, R. (1981). The cognitive representation of an event involving human motion. Cognitive Psychology, 13, 391-406.

Marr, D. (1982). Vision. Freeman.

Marr, D., & Vaina, L. (1982). Representation and recognition of the movements of shapes. Proceedings of the Royal Society of London, 214, 501-524.

McNeill, D. (1996). Language as gesture (gesture as language). In Proceedings of the Workshop on the Integration of Gesture in Language and Speech, Wilmington, DE.

McNeill, D. (1992). Hand and mind: what gestures reveal about thought. University of Chicago Press.

McNeill, D & Levy, E. (1982). Conceptual representations in language and gesture. In R. J. Jarvella & W. Klein (eds.), Speech, Place and Action, 271-295.

Miller, G.A. & Johnson-Laird, P. N. (1976). Language and Perception. Harvard University Press.

Murakami, K. & Taguchi, H. (1991). Gesture recognition using recurrent neural networks. CHI '91 Proceedings. ACM Press.

Pelachaud, C., Badler N., & Steedman. (1996). Generating facial expressions for speech. Cognitive Science, 20(1).

Peterson, P. (1985). Causation, agency, and natural actions. In Chicago Linguistic Society, 21st Region and Parasession on Causative and Agentivity, Department of Linguistics, University of Chicago.

Prevost, S. (1996). An information structural approach to spoken language generation. Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics.

Quek, F. (1994). Toward a vision-based hand gesture interface. In Proceedings of the Virtual Reality System Technology Conference, Singapore.

Rime, B & Schiaraturea, L. (1991). Gesture and speech. In R.S. Feldman & B. Rime (eds.), Fundamentals of Nonverbal Behavior. Cambridge University Press.

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

Saint, P. & Viegas, E. (1995). An introduction to lexical semantics from a linguistic and a psycholinguistic perspective. In P. Saint & E. Viegas (eds.), Computational Lexical Semantics. Cambridge University Press.

Short, J., Williams, E. & Christie, B. (1976). The social psychology of telecommunications. Wiley.

Starner T. & Pentland A. (1995). Visual recognition of American Sign Language using Hidden Markov Models. International Workshop on Automatic Face and Gesture Recognition. Zurich, Switzerland.

Talmy, L. (1985). Lexicalization patterns: semantic structure in lexical form. In T. Shopen (ed.), Grammatical categories and the lexicon, volume 3 of Language typology and syntactic description. Cambridge University Press.

Talmy, L. (1988). Force dynamics in language and thought. Cognitive Science 12, 49-100.

Thompson, L. & Massaro, D. (1986). Evaluation and integration of speech and pointing gestures during referential understanding. Journal of Experimental Child Psychology, 42: 144-168.

Thorisson, K. (1996). Communicative humanoids: a computational model of psychosocial dialogue skills. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge.

Torres, O., Cassell, J. & Prevost, S. (1997). Modeling gaze behavior as a function of discourse structure. First International Workshop on Human-Computer Conversation. Bellagio, Italy.

Tuite, K. (1993). The production of gesture. Semiotica, 93(1/2).

Vaananen & Bohm, K. (1993). Gesture-driven interaction as a human factor in virtual environments - an approach with neural networks. In R.A. Earnshaw, M.A. Gigante & H. Jones (eds.), Virtual Reality Systems. Academic Press.

Walker M. (1993). Informational redundancy and resource bounds in dialogue. Ph.D thesis, University of Pennsylvania, Philadelphia.

Williams, E. (1977). Experimental comparisons of face-to-face and mediated communication: A review. Psychological Bulletin, 84, 963-976.