Distribution of Semantic Features Across Speech & Gesture

by Humans and Machines

Justine Cassell & Scott Prevost

MIT Media Lab

20 Ames Street

Cambridge, MA 02139

justine@media.mit.edu, prevost@media.mit.edu

Introduction

Participants in face-to-face dialogue have available to them information from a variety of modalities that can help them to understand what is being communicated by a speaker. While much of the information is conveyed by the speaker's choice of words, his/her intonational patterns, facial expressions and gestures also reflect the semantic and pragmatic content of the intended message. In many cases, different modalities serve to reinforce one another, as when intonation contours serve to mark the most important word in an utterance, or when a speaker aligns the most effortful part of gestures with intonational prominences (Kendon, 1972). In other cases, semantic and pragmatic attributes of the message are distributed across the modalities such that the full communicative intentions of the speaker are interpreted by combining linguistic and para-linguistic information. For example, a deictic gesture accompanying the spoken words "that folder" may substitute for an expression that encodes all of the necessary information in the speech channel, such as "the folder on top of the stack to the left of my computer."

Deictic gestures may provide the canonical example of the distribution of semantic information across the speech and gestural modalities but iconic gestures also demonstrate this propensity. Most discussed in the literature is the fact that gesture can represent the point of view of the speaker when this is not necessarily conveyed by speech (Cassell & McNeill, 1991). An iconic gesture can represent the speaker's point of view as observer of the action, such as when the hand represents a rabbit hopping along across the field of vision of the speaker while the speaker says "I saw him hop along". An iconic gesture can also represent the speaker's point of view as participant in the action, such as when the hand represents a hand with a crooked finger beckoning someone to come closer, while the speaker says "The woman beckoned to her friend". However, information may also be distributed across modalities at the level of lexical items. For example, one might imagine the expression "she walked to the park" being replaced by the expression "she went to the park" with an accompanying walking gesture (i.e. two fingers pointed towards the ground moving back and forth in opposite directions).

In cases where a word exists that appears to describe the situation (such as "walk" in the above example), why does a speaker choose to use a less informative word (such as "go") and to convey the remaining semantic features by way of gesture? When a word, or semantic function isn't common in the language (such as the concept of the endpoint of an action in English), when does a speaker choose to represent the concept anyway, by way of gesture?

We approach these questions from the point of view of building communicating humanoid agents that can interact with humans -- that can, therefore, understand and produce information conveyed by the modalities of speech, intonation, facial expression and hand gesture. In order for computer systems to fully understand messages conveyed in such a manner, they must be able to collect information from a variety of channels and integrate it into a combined "meaning." While this is certainly no easy proposition, the reverse task is perhaps even more daunting. In order to *generate* appropriate multi-modal output, including speech with proper intonation and gesture, the system must be able to make decisions about how and when to distribute information across channels. In previous work, we built a system (Cassell et al, 1994) that is able to decide where to generate gestures with respect to information structure and intonation, and what kinds of gestures to generate (iconics, metaphorics, beats, deictics). Currently we are working on a system that will decide the form of particular gestures. This task is similar to lexical selection in text generation, where, for example, the system might choose to say "soundly defeated" rather than "clobbered" in the sentence "the President clobbered his opponent" (Elhadad, McKeown & Robin, 1996).

In this paper, we present data from a preliminary experiment designed to collect information on the form of gestures with respect to the meaning of speech. We then present an architecture that allows us to automatically generate the form of gestures along with speech with intonation. Although certainly one of our goals is to build a system capable of sustaining interaction with a human user, another of our goals is to model human behavior, and so we try at each stage to build a system based on our own research, and the research of others, concerning human behavior. Thus, the generation is carried out in such a way that one single underlying representation is responsible for the generation of discourse-structure-sensitive intonation, lexical choice, and the form of gestures. At the sentence planning stage, each of those modalities can influence the others so that we find the form of gestures having an effect on intonational prominence. It should be noted that, in the spirit of a workshop paper, we have left obvious the ragged edges in our ongoing work, hoping to thereby elicit feedback from other participants.

Background

A growing body of evidence shows that people unwittingly produce gestures along with speech in many communicative situations. These gestures elaborate upon and enhance the content of accompanying speech (McNeill, 1992; Kendon, 1972), often giving clues to the underlying thematic organization of the discourse or the speaker’s perspective on events. Gestures have also been shown to identify underlying reasoning processes that the speaker did not or could not articulate (Church and Goldin-Meadow, 1986).

Do gestures play any role in human-human communication? We know that gestures are produced in situations where there is no listener, or the listener cannot see the speaker's hands (Rimé, 1982), although more gestures may be produced when an addressee is present (Cohen, 1977; Cohen & Harrison, 1973). But when speech is ambiguous (Thompson & Massaro, 1986) or in a speech situation with some noise (Rogers, 1978), listeners do rely on gestural cues (and, the higher the noise-to-signal ratio, the more facilitation by gesture). And, when adults are asked to assess a child's knowledge, they are able to use information that is conveyed in the child's gesture (and that is not the same as that conveyed by the child's speech) to make that assessment (Goldin-Meadow, Wein & Chang, 1992; Alibali, Flevares & Goldin-Meadow, 1994). Finally, when people are exposed to gestures and speech that convey slightly different information, whether additive or contradictory, they treat the information conveyed by gesture on an equal footing with that conveyed by speech, ultimately seeming to build one single representation out of information conveyed in two modalities (Cassell, McNeill & McCullough, in press).

We suspect that hand gestures must be integral to communication when we examine their temporal relationship to other communicative phenomena. Hand gestures co-occur with their semantically parallel linguistic units, although in cases of hesitations, or syntactically complex speech, it is the gesture which appears first (McNeill, 1992). At the most local level, individual gestures and words are synchronized in time so that the 'stroke' (most energetic part of the gesture) occurs either with or just before the intonationally most prominent syllable of the accompanying speech segment (Kendon, 1980; McNeill, 1992). At the most global level, we find that the hands of the speaker come to rest at the end of a speaking turn, before the next speaker begins his/her turn. At the intermediate level, the phenomenon of co-articulation of gestural units is found, whereby gestures are performed rapidly, or their production is stretched out over time, so as to synchronize with preceeding and following gestures, and the speech these gestures accompany.

Taken together, these findings lead us to believe that speakers distribute their communicative intention across different modalities, and listeners integrate the information that they receive from the different modalities into one common understanding of the speaker's communicative intention (e.g. Bolt, 1987). Researchers in the human interface community have begun to attend to findings of just this sort, and there is an increasing interest in *multimodal interfaces* that understand speech, gesture, and facial expression. Our own research attempts to go one step further -- we believe that computers should not simply attempt to understand humans, they should generate human-like communicative behavior in response. We design communicative humanoid agents -- animated human figures with faces and hands, and that can produce speech, intonation and appropriately timed gestures and regulatory facial movements. In one previous system (Cassell et al, 1994), we automatically generated the placement of gestures in the stream of speech by using the timing of intonation -- the stroke of gestures co-occured with the pitch peak in intonation. We generated the distribution of gestures in the discourse by using the information structure of the discourse -- gestures co-occured with *rhematic* or new information. We generated the type of gestures that occured by using the nature of the concepts being expressed -- concepts with concrete existence were represented by iconics, concepts commonly conveyed by a metaphor were represented by metaphorics and so forth. In this way, we showed that a computational theory of gesture generation was possible: the occurrence of gestures could be predicted, and gestures and speech could be generated from one common underlying semantic representation.

Two important issues were brought out by the implementation. First, we realized that while a discourse framework could specify type of gesture and placement of gesture, we would need a semantic framework to generate the *form* of particular gestures. In this system we were obliged to choose gestural forms from a dictionary of gestures. In this paper we describe how such a provisional solution may be bypassed, using lexical semantics. Secondly, we realized in watching the animation that *too many* nonverbal behaviors were being generated -- the impression was of a bank teller talking to a foreigner, and trying to enhance his speech with supplementary nonverbal cues. This problem arose because each nonverbal behavior was generated independently, on the basis of its association with discourse and turn-taking structure and timed by intonation, but without reference to the other nonverbal phenomena present in the same clause. Here we discuss a model of the distribution of content across speech and gesture; in other research we are also including facial conversational regulators (Prevost & Pelachaud, forthcoming; Thorisson & Cassell, 1996).

Why Distribution of Semantic Features

Although the distribution of gestures in the discourse, and the choice of type of gestures were generated automatically in "Animated Conversation" (Cassell et al. 1994), we relied on a provisional and unsatisfactory method for generating the form of particular gestures and for deciding what information gestures would convey. A gesture "dictionary" was accessed, which provided particular gesture forms for particular concepts in the domain of the dialogue. This provisional solution produced pre-scripted gestural forms, and gestures redundant with the speech they accompanied, rather than complementary or non-redundant information. So, for example, the gesture in Figure 1 was produced by accessing the gesture dictionary once the concept of "writing a check" had been generated by the discourse planner.

Figure 1: "You can write the check"

Thus the form would remain constant from one mention of the entity to the next, and the gesture would not provide much additional information to the speech "you can write the check". Imagine instead that the agent says "you can do it for fifty dollars" and produces the ‘write-the-check’ gesture illustrated in Figure 1 above. In this case, the dialogue generation module would have sent a "pro-verb" to speech, but filled in the necessary information in gesture.

In other words, we are interested in the issue of *lexical choice,* or how one word *or gesture* is chosen over another. Why do we say "I hightailed it out of the room" rather than "I left the room?" Why do we choose to say "Justine walked to the conference" one day, and another day "Justine went to the conference on foot?" In the first sentence the manner of locomotion is conveyed in the verb, and in the second sentence, the manner of locomotion is conveyed in the prepositional phrase. When we take seriously the idea put forth by McNeill (1992, inter alia) and others that gesture and speech arise from one single underlying meaning structure, then we must add gesture to the lexical choice equation, and wonder how one chooses which meaning features are chosen to be expressed *and* which are expressed in the words and which in the gesture. Additionally, one must wonder why in some cases semantic features are redundantly expressed in speech and gesture, and in other cases non-redundantly -- gesture or speech expressing an aspect of an idea that the other doesn’t convey.

Distribution of Features and Redundancy

There are really two different questions to be addressed with respect to the distribution of semantic features across modalities: (1) is it ever the case that semantic features are distributed across speech and gesture so that some of meaning is conveyed by the gesture, and some is conveyed by the speech? (2) if so, under what circumstances are semantic features distributed so that gestures and speech convey non-redundant information, and under what circumstances are semantic features distributed so that gestures and speech convey redundant information.

Distribution of Semantic Features Across Modalities?

We begin by addressing the question of whether semantic features are always redundant across speech and gesture. Initially we hypothesized that in the domain of manner-of-motion verbs we would be likely to find many non-redundant gestures. That is, we hypothesized that speakers would sometimes express the fact of motion by way of speech (e.g. "he went") and the manner of motion by way of gesture (e.g. a gesture representing walking). Our evidence for this hypothesis came from three sources: (1) from the fact that so many "pro-verbs" exist in English for motion -- "he goes . . .", "he did . . .", "he’s like . . .", and so we imagined that the actual manner of motion might be conveyed by gesture; (2) from anecdotal evidence of the many manner-of-motion gestures viewed during a decade of watching people tell the story of a particularly motion-oriented cartoon; and (3) from the results of an experiment designed to examine the kinds of gestural information taken up by listeners in the course of everyday listening. This experiment showed that listeners incorporate the manner-of-motion gestures they see into their subsequent understanding of what they have heard (Cassell, McNeill & McCullough, in press). We suspected, then, that verbs of motion might be a good test case in terms of the distribution of semantic features across modalities. In particular, we decided to look at the semantic features of ‘manner’, ‘path’, ‘telicity’ (whether a motion has an endpoint or goal), ‘speed’, and ‘aspect’ (inherent iterativity or duration) within verb phrases and gestures describing the movement of volitional agents (Coyotes, and Road Runners).

In a preliminary experiment designed to examine the association of manner-of-motion verbs and gestures, we showed a segment of a Road Runner cartoon to 6 people who told the story to 6 naive listeners. We then examined the semantic features represented in motion verbs and in gesture. We did indeed find a wide variety of verbs of motion, and a wide variety of semantic features of motion expressed in gesture. We also found a very wide variety of prepositional phrases expressing manner of motion. And we did indeed find distribution of semantic features across speech and gesture so that communicative load is shared among the modalities. For example, one subject said "Road Runner comes down", and with both hands makes the gesture of holding the wheel of a car and driving. In this example, in only the *manner-of-motion* gesture do we see that Road Runner's manner of coming down (the road) is to drive. Another speaker described the Coyote in a hot air balloon releasing an anvil tied to a string by saying "he's going to drop the anvil" while he made a gesture of untying a string. The *manner-of-motion* gesture in this case is, in fact, difficult to understand unless one knows that the anvil is dropped by releasing from the string that holds it into the balloon. In fact, examples of other semantic features were even more common than the manner features that were just given, as shown below.

Out of 90 total gestures, the distribution of semantic features was as follows (note that each gesture could display more than one semantic feature).

	path	speed	telicity	manner	aspect
Total #	69	22	5	31	2
Non-redundant	30	15	1	17	0

Thus it would appear that roughly half of the semantic features occured in contexts that were redundant with speech, and half occured in contexts that were non-redundant with speech.

After having read McNeill's contribution to the current volume, however, we re-evaluated our analysis of the manner feature, and came to believe that our coding of manner was conflated with several other variables. First of all, manner features tended to occur with "pro-verbs" or verb+onomatopoiea, such as "the road runner goes pschew". In this case, almost every semantic feature in the gesture must be non-redundant because the speech is simply a pro-form indicating that interpretation must rely on information in the context of utterance, much like the demonstrative in "look at *that* folder". Secondly, several of the manner features that were non-redundant conveyed features that were associated with the *lexical item* they accompanied, but did not convey something about the scene the narrator was describing. These *lexicalized* gestures represent some other sense of the words they accompany. For example, one speaker was describing the Coyote standing in a giant sling shot and taking steps backwards to stretch the sling shot, as a way of catapulting himself towards the road runner. The speaker says "he pulls himself back", and with her two hands represents somebody pulling back a rubberband with his hands. Of course, Coyote does not use his hands, but his body to stretch the slingshot, but the verb used is ambiguous as to manner, and the gesture is representing the other sense of the verb. Finally, we began to see a new category emerge, that we called "manner / path ambiguity". An example is the speaker who is describing Coyote running and crashing through a canvas painting of a road that is camouflaging a cliff, and then falling down the cliff and says "Coyote goes through it and falls" while making a gesture that makes a sharp movement to the right and then little circular hovering motions in the air before falling straight down. The gesture is indeed describing a path, through the air in little circles and then down. But this path has a manner name: we call it "hovering in the air". We will return to the consequences of this re-analysis of manner below. Here, however, it suffices to say that we were right to believe that semantic features were indeed distributed across modalities in the domain of motion verbs. This gives us reason to think that we can generate the form of gestures by using a lexical choice-like approach (a semantic frame analysis, as described below).

Redundancy versus Non-Redundancy: the Case of Manner of Motion

The next issue, then, is the distribution of redundant and non-redundant gestures: when do gestures convey complementary information, and when do they emphasize the same concept as speech. The category of manner, which we believed would yield examples of redundant and non-redundant gesture-speech pairs, was not the unitary category we believed. However the problems with this analysis, viewed in the light of McNeill's paper, give us the beginnings of an answer. Why didn't we find many non-redundant manner gestures? McNeill claims that one doesn't find as many non-redundant manner gestures as path because English has too many manner *verbs* that do the work. McNeill also claims that the form of gesture is associated with the "most newsworthy" information, and that path is more frequently noteworthy than manner. We would say, rather, that this is due to genre. Road Runner and Tweetie Bird cartoons do not rely heavily on modes of transportation, making it unlikely that the *manner* of motion be the most newsworthy information in a unit. The bottom line, however, is that *overmarking* or redundancy across modalities appears to mark rhematic or newsworthy or less predictable information. These two facts may account for why we found as few examples as we did. What accounts for the distribution of where we did find them? We did find manner gestures conveying complementary non-redundant information in the context of "demonstrative onomatopoeia" -- sentences such as "he went whoosh". Note that the complex gestures that accompanied these performances make it likely that the speaker chose this lexical instantiation because no adequate verb existed in English to convey the speaker's intention. We also found complementary non-redundant manner verbs in the context of verbs of action that did not have to do with manner of motion. Rather, for the most part these verbs described some activity, and the verbs that accompanied them were all performed with a character viewpoint (Cassell & McNeill, 1991). An example comes from a speaker describing Coyote spreading cement with a trowel. She says "he's like putting this on the road" and represents spreading with a trowel with her hands. The preponderance of non-redundant manner gestures in this context, we believe, is an artifact of the fact that the trajectory of the gesture is not likely to be taken as an expression of path. That is, although we have adopted the suggestions of McNeill in this volume as to the function of redundant gestures, we do not agree that path is so frequently represented because it is so frequently part of what makes an opposition significant. Rather, we find that -- given that all gestures are made up of a handshape, a location in space, and a trajectory -- path is the *easiest* semantic feature for gesture to represent, since the trajectory of a gesture will always appear to mark a path, when accompanying speech is referring to motion.

Computational Architecture

Based on our hypothesis--that various features of motion-verb phrases are distributed across communicative modalities--and our analysis of the experimental data described above, we propose a computational architecture for generating face-to-face spoken language with appropriate distribution of semantic load across speech, intonation and gesture. This architecture refines and expands the architecture presented in Prevost (1996), which generates monologic descriptions of objects represented in a small knowledge base. By automatically annotating propositions with theme/rheme distinctions and applying an algorithm to identify contrastive relationships, this system was able to produce synthesized speech with contextually appropriate intonation. Since the system presented in Cassell et al. (1994) generates gestures using the same types of discourse relationships (theme/rheme, given/new, contrast, etc.), the integration of similar discourse-sensitive gesture production into the monologue generator is relatively straightforward. As we described earlier, however, Cassell et al. (1994) succeeded in predicting the placement and type of gesture, but not the *form* of gesture. Also, that system was unable to account for gestures which convey information that is non-redundant with the information given in the speech stream. Our goal in this section is to describe an extension of that architecture that allows for the production of redundant and non-redundant gestures relating to motion verbs.

A key component of the proposed architecture is a semantic representation scheme that encodes the proper level of abstraction for concepts involving motion so that features such as manner, path, telicity, speed and aspect can be independently applied to the various modalities at hand. So, for example, given a hypothetical system with multi-modal input, the gesture recognizer might identify a path of motion while the speech recognizer might identify the manner, or vice versa. Given our knowledge of the relationship between intonational phrasing and gesture placement (Kendon 1972), such a system would be able to unify the two inputs into a single frame representing the meaning of the combined speech and gesture, as illustrated in Figure 2.

Gesture Frame Speech Frame Semantic Frame

"comes down" "comes down"

Path: **[[union]]** Path: *down* Path: *down*

Manner: *drive* Manner: Manner: *drive*

Speed: Speed: Speed:

Telicity: Telicity: Telicity:

. . . .

Figure 2: Unification of Frames

That is, we're looking for, on the one hand, a*frame semantics* for motion verbs that will allow us to specify the features within the frame that make up the sense of the verb, and will allow us to disassociate those features from the medium in which they are realized. And on the other hand, we need a discourse framework that will allow us to mark the conditions under which information receives "overmarking" or expression by two modalities. Both of these must fit into one unified framework for the generation of speech and gesture.

The proposed architecture, shown in Figure 3 is based on the monologic generator described in Prevost (1996), which was originally conceived to produce descriptions of objects. The task of natural language generation is divided into three stages: content planning, in which high-level goals are satisfied and discourse structure is determined (the "discourse framework"), sentence planning (wherein one finds the frame semantics), in which high-level abstract semantic representations are mapped onto lexicalized representations (Rambow and Korelsky 1992, Reiter and Mellish 1992, Meteer 1991) and surface generation, in which the lexicalized representations are converted into strings of words.

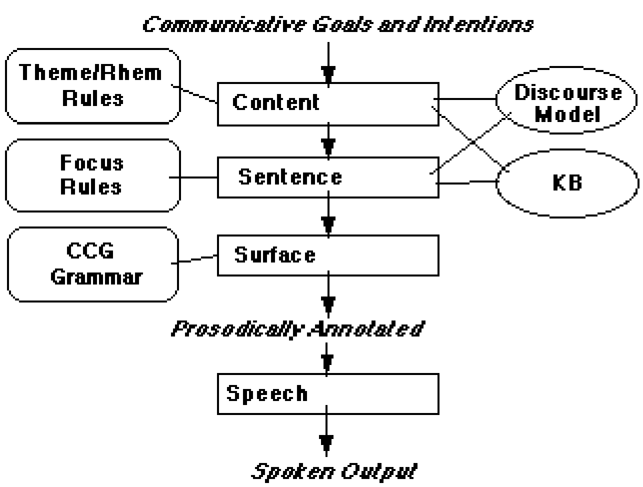

Figure 3: System Architecture

In Prevost (1996), the selection and organization of propositions is determined in the content planning stage using a hybrid of McKeown's (1985) schemata-based system and rhetorical structure theory (RST) approaches (Hovy 1993, Mann and Thompson 1986). It is at this stage that the discourse structure is determined and propositions are divided into their thematic and rhematic constituents. Based on previous work in Cassell (1994), gestures are placed so as to co-occur with the rhematic material. Consequently, the content generator determines the alignment of gestures with the high-level propositions and their information structure representations. One key difference between the original content planner and the one required here is related to the task at hand. Whereas Prevost (1996) was concerned with descriptions of objects, we are now concerned with descriptions of events. Consequently the existing rules that identify rhetorical relationships among properties must be augmented with new rules that convey the types of rhetorical relationships generally found in event descriptions, such as causation and temporal sequence.

The second phase of generation, sentence planning, is responsible for converting the high-level propositions from the content planner into representations that more fully constrain the possible sentential realizations. In general, this stage can be viewed as the bridge between the primarily language-independent content planning and the highly language-dependent syntactic rules. Language specific issues that require access to a global discourse model, such as building referring expressions (Dale and Haddock 1991) and selecting among lexical alternatives, are often handled in the "sentence planning" phase of generation.

Since the distribution of semantic features across communicative modalities is certainly language specific (as argued by McNeill, this volume), we take the determination of such distributions to be in the domain of the sentence planner. So, just as the issue of lexical choice, where semantic features are distributed across parts of speech, is handled at the sentence planning stage, so should the issue of modality choice be handled at this stage as well. Our sentence planner might therefore encode a rule for English that always chooses to represent the path of motion in gesture and the manner in speech. For Spanish, we might encode a rule that opts to place the manner feature in gesture. So, while the content planning phase is responsible for gesture placement, the sentence planning phase is responsible for selecting the features to be conveyed by the gesture.

One further point concerning the sentence planning is worth noting. Prevost (1996) argues that the determination of focus (and hence pitch accent placement) within thematic and rhematic constituents should be handled by the sentence planner. Based on this observation and the mapping of tri-phasic gestures onto intonational tunes described in Cassell (1994), we can also assert that the alignment of the three gesture phases with the intonation contour occurs at this level as well. This aspect of our architecture has a strong effect on the interaction between speech and gesture in generation: the choice of gestures and choice of speech form interact such that gesture will actually affect where stress is placed in the utterance. For example, if a sentence such as "Road Runner zipped over Coyote" is planned then, depending on the gesture chosen, as well as the underlying representation, primary stress will be differently assigned. If the gesture chosen represents driving, then primary stress will fall on "zipped" (as the reader can see by reading the sentence out loud, it is difficult to imagine performing the gesture along with "over", or stressing the word "over" if the gesture co-occurs with "zipped"). If, on the other hand, the gesture chosen simply represents motion from point A to point B, then primary stress might fall on "zipped" or on "over" depending which of these terms is focused (or contrastive) in the context of the text.

The final stage of generation consists of building a surface form (words, intonation and gesture) from the output of the sentence planner. In Prevost (1996), a Combinatory Categorial Grammar generator is used to translate lexicalized logical forms, which include information structure and focal articulations, into strings of words with intonational markings. In the new architecture proposed here, the generator produces gestural forms by realizing the appropriate semantic features as specified by the sentence planner. So, whereas in previous work (Cassell 1994), the form of a gesture was rigid, we now allow a given concept to be represented by a variety of forms based on the output of the sentence planner. For example, a concept like "driving" might be realized by the verb "drive" with an accompanying path gesture, or by the less-specific motion-verb "zip" with an accompaning driving (two hands on the wheel) gesture. The rules instatiated by the system are the following:

* In the unmarked case, distribute semantic features across speech and gesture.

That is, look first at what is perceptually salient in the scene (Herzog & Wazinski, 1994), and then look at the lexicon of the language for what is likely and able to be marked in language, and what in gesture (see Kita, 1993 and McNeill, this volume) among the salient features.

* In the marked case, *overmark* , or add redundance to the expression of concepts by conveying them in speech and gesture.

That is, when something is rhematic, or contrastive or focused, then overmark it. Likewise, overmark it if the item participates in a lexical collocation (a continuation of items that fit together -- drive, run, walk, bike). Figure 4 shows automatically annotated and intoned output.

The coyote is flying in a balloon with a large anvil.

He drifts over the roadrunner and drops the anvil.

The balloon deflates and the coyote falls to the ground.

The anvil drops onto him.

And then, the roadrunner zooms over him.

Figure 4: Automatically Generated Output

Conclusions

With this addition of a sentence planning layer that accounts for lexical choice and gestural form to our generation engine, we come closer to our goal of generating speech and gesture from scratch. We have still to implement this layer in sufficient detail to handle complete event descriptions and complex gestures. Also still to be determined is how the actual surface form of the gesture will be generated (e.g. will "drive" be represented by a closed hand or an open hand, by a single-handed gesture or a two-handed gesture). But we have shown how data from natural spontaneous gesture can be used to write predictive rules that allow us to translate human behavior into machine behavior. References

Alibali, M.W., Flevares, L. & Goldin-Meadow, S. (1994). Going beyond what children say to assess their knowledge. Manuscript, Department of Psychology, University of Chicago.

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.

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

Cassell, J., McNeill, D. & McCullough, K.E. (in press). Speech-gesture mismatches: evidence for one

underlying representation of linguistic & nonlinguistic information. *Cognition*

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 (SIGGRAPH proceedings).*

Cassell, J., Stone, M., Douville, B., Prevost, S., Achorn, B., Steedman, M., Badler, N., and Pelachaud, C.

(1994b). Modeling the interaction between speech and gesture. *Proceedings of the Sixteenth Conference of the Cognitive Science Society* (August, 1994: Georgia)

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 functions of hand illustrators. *Journal of Communication*, 27(4): 54-63.

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.

Dale, R. & Haddock, N. (1991). Content determination in the generation of referring expressions. *Computational*

Intelligence, 7(4): 252-265.

Elhadad, M., McKeown, K. & Robin, J. (1996). Floating constraints in lexical choice. *Computational Linguistics*.

Goldin-Meadow, S., Wein, D. & Chang, C. (1992). Assessing knowledge through gesture: using

children's hands to read their minds. *Cognition and Instruction*, 9(3): 201-219.

Herzog, G. & Wazinski, P. (1994). VIsual TRAnslator: linking perceptions and natural language descriptions.

Artificial Intelligence Review, 8: 175-187.

Hovy, E. (1993). Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63: 341-385.

Kendon, A. (1980). Gesticulation and speech: two aspects of the process. In M.R. Key (ed.), *The*

Relation Between Verbal and Nonverbal Communication. Mouton.

Kendon, A. (1972). Some relationships between body motion and speech. In A.W. Siegman & B. Pope

(eds.), *Studies in Dyadic Communication*. New York: Pergamon Press.

Kita, S. (1993). Language and thought interface: a study of spontaneous gestures and Japanese mimetics. Ph.D.

dissertation, Department. of Psychology (Cognition and Communication) and Department of Linguistics, University of Chicago.

Mann, W. & Thompson, S. (1986). Rhetorical structure theory: description and construction of text structures. In

G. Kempen (ed.), *Natural Language Generation: New results in Artifi cial Intelligence, Psychology and Linguistics*. Boston: Kluwer Academic Publishers.

McKeown, K. (1985). *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought.* Chicago: University of

Chicago Press.

Meteer, M. (1991). Bridging the generation gap between text planning and linguistic realization. *Computational Intelligence*, 7(4): 296-304.

Prevost, S. (1996). An information structural approach to monologue generation. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics* (June, 1996: Santa Cruz).

Prevost, S. & Pelachaud, C. (forthcoming). *Talking Heads: Synthetic Faces and Spoken Language.* Cambridge, MA: MIT Press.

Rambow, O. & Korelsky, T. (1992). Applied text generation. *Proceedings of the Third Conference on Applied Natural Language Processing (ANLP '92)*: 40-47.

Reiter, E. & Mellish, C. (1992). Using classification to generate text. *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*: 265-272.

Rimé, B. (1982). The elimination of visible behavior from social interactions: effects of verbal, nonverbal and interpersonal variables*. European Journal of Social Psychology*, 12: 113-129.

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

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

Thorisson, K. & Cassell, J. (1996). Why put an agent in a body: the importance of communicative feedback in human-humanoid dialogue. Lifelike Computer Characters '96 (Snowbird, October, 1996).