Content in Context:
Generating Language and Iconic Gesture without a Gestionary

Paul Tepper
Northwestern University
2240 Campus Drive
Evanston, IL 60208
+1 847 491 4624
ptepper@northwestern.edu

Stefan Kopp
Northwestern University
2240 Campus Drive
Evanston, IL 60208
+1 847 467 4662
skopp@techfak.uni-bielefeld.de

Justine Cassell
Northwestern University
2240 Campus Drive
Evanston, IL 60208
+1 847 491 3534
justine@northwestern.edu

ABSTRACT

When expressing information about spatial domains, humans frequently accompany their speech with iconic gestures that depict spatial, imagistic features. For example, when giving directions, it is common to see people indicating the shape of buildings, and their spatial relationship to one another, as well as the outline of the route to be taken by the listener, and these gestures can be essential to understanding the directions. Based on results from an ongoing study on gesture and language during direction-giving, we propose a method for the generation of coordinated language and novel iconic gestures based on a common representation of context and domain knowledge. This method exploits a framework for linking imagistic semantic features to discrete morphological features (handshapes, trajectories, etc.) in gesture. The model we present is preliminary and currently under development. This paper summarizes our approach and poses new questions in light of this work.

1. INTRODUCTION

In this paper we describe new research on the planning and realization of paired natural language and gesture for embodied conversational agents. Some gestures (so-called *emblems*) constantly exhibit the same, characteristic form (e.g., a "V" formed with the forefinger and the middle finger as a symbol for peace). We focus here on *iconic* gestures that communicate in virtue of their resemblance to aspects of concepts or (mental) images. Gestures do not always convey the same meaning in the way that words do – in fact, there is a many-many relationship between gestural form1 and concepts, and so in generating gesture, we cannot rely on a fixed lexicon of gestures, or "gestionary", but have to search for ways to derive the form of the gesture on the fly. As in our previous work, we rely on the study of spontaneous gesture to inform us about the relationships between spontaneous hand gestures and language, and we rely on models of natural language generation to inspire our computational architectures. Unlike our previous work, however, here we work towards a formalization of both the imagistic and linguistic components of people's cognitive representations of domain knowledge. This involves modeling the generation process in a way that allows the same representations and communicative intentions to be pursued across a range of communicative modalities and, ultimately, in parallel ways in both input and output.

A balanced model of action and perception requires multimodal input integration and understanding to rely on the same representations as output planning and generation. In this view, while the system thus contains several distinct subsystems, there is a uniform representation of meaning and the evolving discourse context both for input and output [25]. In our past work, we have been committed to the view that a uniform representation throughout the architecture simplifies and facilitates the construction of a balanced and complete representation of the evolving context of conversation.

The REA system, an ECA which generates context-appropriate, coordinated language and gesture [6], is based on empirical evidence [4],[28] that communicative content can be regarded in terms of semantic components, and that different combinations of verbal and gestural elements can be associated with different strategies to distribute these components across the modalities. Yet, the uniform representation of information only extended as far as the first stage of planning in output, since gesture form was chosen from a library of pre-determined gestures. The MACK system includes a repertoire of capabilities in non-verbal behavior perception and generation [8]. These capabilities serve as the foundation for an implemented model of face-to-face grounding for direction-giving [21]. The system continuously observes the user's eye gaze and head nods as well as verbal backchannel cues. Using these cues, the system keeps track of the user's understanding and accordingly decides whether to elaborate a given dialogue act further, or to go on to the next one [1]. In MACK, however, nonverbal behaviors can only superficially be integrated into representations of discourse history.

Our current project, NUMACK, picks up where MACK and REA left off. NUMACK, an interactive direction-giving kiosk with an embodied conversational agent, answers questions about locations and buildings on Northwestern University's campus and provide directions to each. First, using the information state approach to dialogue management [26], we keep a common representation of discourse history including grounded and ungrounded facts available as a resource that informs choices throughout generation [20]. Second, and more importantly here, we focus on generation of natural, novel iconic gestures to accompany language, from a shared representation of meaning.

The following sections describe our current model in detail, which extends the microplanning stage of a common natural language generation (NLG) architecture to generate appropriate, novel (i.e. not predefined), iconic gestures that share the communicative work with speech. We introduce an intermediate level for representing the visual, spatial, and image-evoking, or *imagistic*, information that is expressed in iconic gestures. We propose a framework for linking these imagistic semantic features to discrete morphological features (handshapes, trajectories, etc.) in gesture. Our system is currently under development and only partially realized. Some parts that have been developed in previous projects [6] are being extended and integrated. Other parts are being constructed presently, in parallel with the data collection and analysis that informs them.

2. GENERATING COORDINATED LANGUAGE AND GESTURE

NLG architectures are commonly implemented in a modular, pipeline architecture [22], broken down into three subtasks—*content planning* (also known as text or document planning), *microplanning* and *surface realization* (in that order). In ordinary language, the work done by these three subsystems boils down to, figuring out what to say, figuring out how to say it and finally, saying it, respectively. Generating natural language with gestures (henceforth NLGG) for ECAs will require specialized modules at every stage in the pipeline:

1. **Content Planning.** Selecting domain-specific knowledge (content) to be conveyed and organizing it into a rhetorically structured plan.
2. **Microplanning.** Taking the content plan and recoding it into both coordinated linguistic terms (also known as sentence planning) and gestures (sometimes termed gesture planning [11]).
3. **Surface Realization.** Turning the linguistic structures into morphologically and phonologically-specified speech and intonation, as well as planning gesture motions for a graphical avatar body.

Content Planning. A content planner for NLGG requires a knowledge base with rich representations of domain knowledge. In developing these representations, we pay attention to the *affordances* of gestures as a medium or mode of output [5]. We assume both that gestures are communicative and that some kinds of information are easier to convey in gesture than in spoken language—i.e., information expressible using the depictions possible with hand shapes and motion. The information iconic gestures convey must be visual, spatial and image-evoking, or *imagistic*. In our current project on direction giving, we look primarily at spatial information about locations, actions, or the shape of landmarks.

Microplanning. To construct multimodal utterances, SPUD, a grammar-based microplanner [24], is employed. SPUD iteratively builds utterances using a greedy search algorithm, wherein microplanning is framed as a constraint-satisfaction problem. Constraints are imposed by three input specifications for the system. First, linguistic resources include: lexical entries, which connect words to logical formulae defining the meaning (semantics) and conditions for use in conversation (pragmatics); and syntactic entries, which comprise grammatical structures, or trees in Lexicalized Tree Adjoining Grammar (LTAG), associated with similar pragmatic formulae, as well as sets of words which may “anchor” these trees. Second, a knowledge base consisting of facts about the domain, explicitly labeled with information about their conversational status, e.g. whether the fact is *private* or *shared,* constraining decisions about what information the system must assert and what it can presuppose as information on the common ground [3]. Third, a dialogue manager maintains the continually evolving context in the information state.

Cassell, Stone & Yan [6] used SPUD in the REA system, extending its linguistic resources for gesture as follows. Whole gestures were treated like words, given lexical entries and associated with a set of one or more semantic and pragmatic formulae. A special grammatical structure was used so that a placeholder for a gesture could be inserted directly into the syntactic tree being constructed for the utterance. This device allowed for a simple solution to the problem of temporal synchronization gestures and the words they relate to. However, treating whole gestures as words does not allow for the expression of new content in gestures, as is possible in language, using a finite set of words and a generative grammar for combining the words into new sentences. Still, in principle, an approach similar to that taken by Cassell, Stone & Yan could work for constructing new gestures on the fly. We present such an approach in Section 3.

Surface Realization. The last stage in the NLGG pipeline concerns generating and executing planned communicative behaviors with a graphical avatar’s body and its synthetic speech. For this problem, we build on the previous BEAT system [7] that is able to annotate textual input with nonverbal and paraverbal behaviors—eyebrow raises, eye gaze, head nods, hand gestures, as well as intonation contours—and to schedule those behaviors with respect to synthesized text output. In our current approach, we employ BEAT’s rule-based components for selecting additional communicative behaviors as well as for scheduling verbal and nonverbal behaviors, but this time on the basis of underlying representations which are provided by the microplanner. However, as we cannot rely on canned gesture animations, we add a module for calculating the required animations on the fly. Most work on gesture animation for ECAs relies on using static libraries of predefined motion elements (e.g. [7]) and applying procedural animation to adjust [9] or to combine them (e.g., [16]). We tackled this problem in the previous MAX system [13], which uses a generation model that creates all verbal and gestural behaviors from formal specifications of their overt form. In particular, this system comprises a hierarchical model for calculating and controlling upper-limb movements of the avatar's skeleton in real time, which allows for flexibility with respect to the producible forms of gesture, and a fine adaptation to temporal constraints as imposed by cross-modal synchrony. This model will be integrated as an additional behavior realization module at the end of the BEAT pipeline, i.e. after the scheduling step. It will take the gesture form definition originating from microplanning and timing constraints set up during scheduling as input, and turn them into applicable motor programs to drive NUMACK's body.

3. GESTURE AS A MICROPLANNING PROBLEM

There are some basic differences between the kinds of meanings iconic gestures can have, and the kinds of meanings, or lexical semantics, posited for words or morphemes. Words are arbitrarily linked to the concepts they represent, gestures are not. Iconic gestures communicate in virtue of their resemblance to the concepts they represent, words for the most part do not. Communicative acts, and the language that comprises them (words, sentences, discourse, etc.) have intended meanings. Many words are *polysemous*, or have multiple meanings, and are therefore ambiguous without the proper context to clarify their intended meaning. But, while words may be ambiguous without context, decontextualized gestures are necessarily vague, with an infinite number of possible interpretations. Even a well-described, specific gesture, e.g. holding one's right hand flat, palm facing left, fingers pointed away from the body, and moving one's hand horizontally away from the body, has a potentially infinite number of interpretations in isolation.

From the point of view of the observer or listener, decontextualized gestures are vague. That is, without context, they don't unambiguously pick out specific or concrete entities, like objects or events. Observers view gestures only as consisting of handshapes and movements in space. But, recent empirical evidence [12],[23] suggests that there are *patterns* in the form and function of iconic gestures with respect to expressing spatial information and communicating meaning more generally. For example, one can find certain consistencies in the mapping from morphological form features to shape features and relations. A flat handshape resembles a flat two-dimensional plane, and a horizontal movement resembles a horizontal extent or axis. Unlike spoken language, however, in gesture multiple form features may be combined to express multiple spatial aspects (extent *and* shape, for example) at once. Likewise, depictions of complex spatial structures may be broken into features that are additively built up by successive gestures. The fact that one same spatial structure is referred to (a winding road, for example) is signaled by spatial coherence; that is, the gestures employ the same viewpoint, size scale, and frame of reference, as indicated by a constancy of handshape, trajectory and position in space. Sometimes, the frame of reference (such as the winding road) is explicitly anchored in gesture space by one hand, and then held throughout while the other hand describes describing additional landmarks at appropriate relative locations.

Based on this empirical evidence, we propose that a system for formalizing the images conveyed by decontextualized gestures should describe shapes, spatial properties and spatial relationships. When such a description of a gesture is then placed in linguistic context, and unified with the semantics for speech, the set of possible interpretations for the gesture becomes so constrained as to make it unambiguous. Therefore, the interpretation of a gesture is crucially dependent on the language it accompanies and the context in which it is articulated. This framework allows us to compose gestures, beyond Cassell, Stone & Yan's REA system [6].

At this point, it is important to address the apparent contradiction of this view – that gestures can be composed out of sub-parts that carry meaning – with the idea that gestures are *global*, or that the meanings of the parts of a gesture are determined by the whole [18]. These two views are not at odds, but instead apply to different levels of abstraction, the imagistic level and the semantic level. In the next section, we walk through an example taken from data that informed the REA system [28]. We use this example to illustrate how gestures can be constructed from vague representations of imagistic properties, and then linked to representations of context and meaning. We will also address the ensuing distinction between the levels of imagistic and semantic analysis.

3.1. Example and Analysis of Communicative Intent

Example (1) was produced by a subject describing a house; (1S) is the speech component of the utterance and (1G) the gesture.

(1) (S): It has a <u>large porch (*pause*) in front.</u>
(G): [*iconic gesture*]

The underlined text indicates the duration of the meaningful phase of the gesture phrase, or the *stroke*, shown in Figure 1, after which the hands retract. Before saying "large" the subject raises his hands up towards the position where he makes his gesture, in the preparation phase of the gesture phrase. Following McNeill [18], we assume that the stroke corresponds to the words with which it temporally co-occurs. In making the gesture, the subject uses both hands; both are in the same shape, close to flat, with the fingers pointed away from body (towards the camera), and slightly curved downward at the end, almost as if in a very loose ASL *C* handshape; thumbs point down

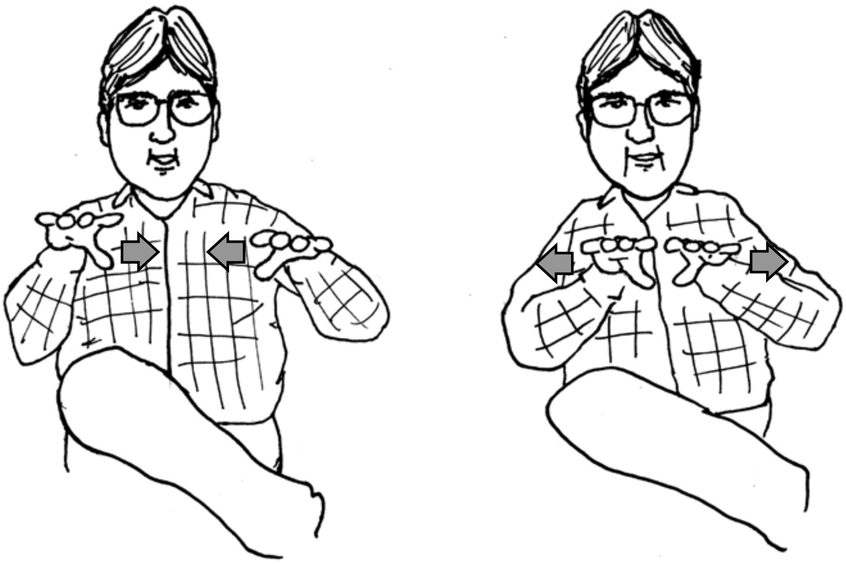

Figure 1. Iconic gesture (1G) accompanying (1S).

and palms face downward; for the trajectory of the motion, both hands moving horizontally away from and back in towards each other several times, tracing a flat, 2-dimensional area, slightly wider than his shoulders. If the hands are seen as more *C* shaped, this might be a rectangular area or cylindrical, 3-dimensional area, given that the slight curvature of the handshape gives some height to the shape traced in the gesture, in addition to the two dimensions of width and depth.

We posit that the meaning of this gesture is that the porch's "largeness" is an extent in the two-dimensional plane parallel to the ground, and that the specific large dimension is the horizontal axis from left to right. The left-to-right axis is a spatial feature of *both* the handshape, which is relatively flat and thus 2-d, *and* the motion, whereas the depth (or the horizontal axis away from the speaker) is only conveyed by the handshape, as there is no motion along this axis. Had the speaker only wanted to convey the one dimension, we might guess that he could have used a different handshape, e.g. tracing the horizontal axis line with his index fingers, with the same motion.[1]

The redundant usage of speech and gesture to convey the same information is typically thought to place special *emphasis* on the shared meaning [4], similar to the way intonation contours can be used to focus attention. In this example, language and gesture seem to convey different, *complementary* information, but as noted, the information conveyed by the two form features in the gesture overlap, or are partly redundant. The handshape conveys two spatial features (depth and width) and motion conveys only one of these features (width). We assume that the overlap, or redundancy in the horizontal feature, indicates that the *salient* aspect of the image being conveyed is the width. While the speaker wants to show the 2D spatial nature of the porch, he also wants the redundancy to emphasize the width, so that the hearer will interpret the *largeness* feature as being associated with this more salient dimension.

We now return to the distinction between levels of imagistic and semantic analysis, introduced in the last section. As this example shows, imagistic descriptions apply to decontextualized gestures, before unification with context. This level describes shape and spatial relationships in terms of structure and composition. Such an approach is common in work on image perception, in computer vision and cognitive psychology [1],[17],[14]. Yet, at the semantics level, the meaning of language and gesture combine such that gestures connect to particular referents, e.g. the *porch*. At this level, the gesture can only be interpreted holistically. Upon interpretation, the meaning of the gesture's parts, e.g. the flat handshape and its trajectory, are determined by the meaning of the whole gesture, e.g. they become the shape, width and orientation of the *porch*.

3.2. Formalizing Meaning in Communicative Intent

Here, as in all natural language generation, *reference* is at the heart of the problem. Reference links communicative acts, in both language and gesture, to the context. For

[1] Intuitively, one might think that the height of the porch was intended as part of the meaning of the gesture due to the somewhat curved shape of the hand. However, the speaker uses the same gesture before in a context where it seems clearer that he is indicating a 2-d surface (*a brick façade*), as opposed to a 3-d surface. Although it is possible he is indicating largeness in three dimensions (depth, width and height), for the purpose of keeping this example simple, we are assuming here that he only intended to refer to two dimensions (depth and width). When a gesture is recurs in a conversation, McNeill [19] posits that it is usually used to evoke the same image, and calls this a *catchment.*

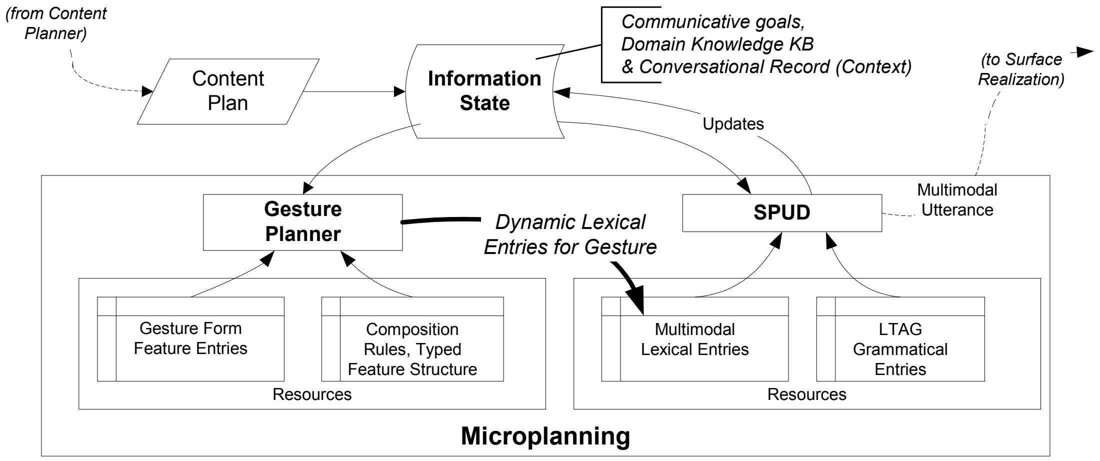
(from Content Planner)
Content Plan
Information State
Communicative goals, Domain Knowledge KB & Conversational Record (Context)
(to Surface Realization)
Updates
Multimodal Utterance
Gesture Planner
Dynamic Lexical Entries for Gesture
SPUD
Gesture Form Feature Entries
Composition Rules, Typed Feature Structure
Resources
Multimodal Lexical Entries
LTAG Grammatical Entries
Resources
Microplanning

Figure 2. Multimodal Microplanning Architecture.

instance, the use of the pronoun "it" in (1S) *presupposes* that the referent house being described is already in the *common ground* [3]. Similarly, "has a large porch" *asserts* the existence of a new referent, the porch; a relationship between the house and the porch ("has"); and a property ascribed to the porch ("large"). Thus reference takes center stage in our computational model of this generation process. Formalizing the semantics of language and gesture allows us to represent the links between the surface forms of utterances and an agent's knowledge of context and the world within which it is situated (domain knowledge).

In formalizing the content planner's input communicative goal of describing the porch, we use a logical formula, *describe*($p1$), where $p1$ is a *discourse referent*, representing an entity, or more specifically, a physical object in the world. Here we use $p1$ to denote the porch. Upon processing this input, the content planner then returns a plan specifying the necessary domain knowledge (or content) required to fulfill the goal of describing the porch.

Based on previous work, we know the kinds of knowledge that SPUD would need to generate a linguistic utterance like (1S). That is, a representation of semantic information to be conveyed by the sentence, in addition the grammar for generating the utterance and lexical entries for each of the words. Since we have already analyzed the meaning of the gesture for (1), in the remainder of this section, we will show how the input knowledge needed to build the communicative intent representation for SPUD to generate (1S) & (1G) can be formalized. We formalize this knowledge in terms of the Prolog-like logic used in SPUD. This provides an elegant interface between the generation of natural language and gesture in microplanning. However, the question of whether this representation is ideal, or even adequate, for the information we must represent, remains open, to be determined by the analysis of the data we are currently collecting and evaluation of the system (cf. Section 4).

To generate (1S), we start with the main verb, "has", and its arguments, the haver, "it," which refers anaphorically to a house being described, and the porch, which we represented earlier using the discourse referent $p1$. Similarly we can denote the house as $h1$, and the having event as $e1$. The existence of the porch is being asserted, as is its property of largeness, but the house must be presupposed, as indicated by the pronoun. Lastly, the preposition "in front" indicates the location of the porch, relative to the house. So we begin with some initial facts, which can be represented like this:

(2) **Presuppose:** $\mathit{house}(h1)$
Assert: $\mathit{porch}(p1) \wedge \mathit{property}(l1, \mathit{large}, p1)$
$\wedge\ \mathit{rel\_loc}(h1, p1, \mathit{in\text{-}front\text{-}of})$

Next, through the gesture, several spatial facts are expressed, pertaining to the orientation, width, and depth of the porch. By themselves, the form features of the gesture each express an abstract property, such as the fact that the porch occupies a two-dimensional area and that it is horizontal. These properties are not quantitative spatial values, as they do not express the extent of the porch in either dimension with any precision or along any obvious quantifiable scale. So it could suffice to represent the meaning of the descriptions required with vague terms:

(3) **Assert**: $\mathit{orientation}(\mathit{otn}1, p1, \mathit{horizontal})$
$\wedge\ \mathit{width}(w1, p1) \wedge \mathit{depth}(d1, p1)$

Note however, that our analysis of the full utterance including the gesture told us that the property of largeness was actually associated with a particular dimension, namely the width. By associated *qualitative* features describing the extent of the dimensions along a qualitative scale, we can distinguish between the *width* and *depth* features, without needing any quantitative information. Since the *width* feature is already associated with the porch, we can simplify the representation by replacing the *property* feature with an extent feature, affiliated specifically with width, the salient dimension, resulting in the following:

(4) **Assert**: *porch*($p1$) ∧ *rel_loc*($h1$,$p1$,*in-front-of*) ∧ *orientation*(*otn*1, $p1$, *horizontal*) ∧ *width*($w1$, $p1$) ∧ *depth*($d1$, $p1$)) ∧ *extent*($w1$, *large*) ∧ *extent*($d1$, *normal*)

This representation also requires the additional definition of a qualitative scale of extent, along which values like *large* and *normal* would fall. Thus, based on this example, we can get an idea of the kinds of semantic formulae we want our content planner to send to SPUD.

3.3. Gesture Planning: Meaning vis-à-vis Context

In a SPUD lexicon, lexical entries associate words with formulae specifying their semantics. These formulae are expressed in terms of discourse anaphora, the open variables mentioned earlier. For example, the meaning of the word "porch" would simply be *porch*(X), so, when SPUD selects this word in generating a sentence like (1S), this referring expression is represented by an inferential link from X to the intended referent $p1$, realized by unification. So, achieving the input goal of asserting a fact like *porch*($p1$) is simply a matter of retrieving a word with the appropriate semantics, *porch*(X), and recording the inferential link to *porch*(*p1*) as part of the communicative intent being planned. Planning novel gestures can be similar —by associating semantic components to choice of particular form features, or gesture morphology.

The SPUD algorithm composes sentences in part by starting from an LTAG initial tree and iteratively filling empty substitution sites in the tree until it has added enough words to achieve the desired communicative goals. Similarly, the GP iteratively fills empty features until a whole gesture is composed. All features are qualitative and discrete, restricting the GP to vague formulations of gesture form, and for any input set, there may be several possible gestures capable of expressing the desired content.

We propose the addition of a new subsystem for *gesture planning* within the microplanning stage, as illustrated in Figure 2, which will be responsible for planning the form of new gestures from a set of one or more input semantic facts. The gesture planner (GP) is itself a microplanner solving the problem of connecting morphological or form components of a gesture to semantics. In microplanning a multimodal utterance, the SPUD algorithm will be extended to call the GP with all the communicative goals it received. The GP's task is to create gestural realizations for a constellation of communicative goals and to deliver these realizations in the form of lexical entries. These lexical entries are added to the linguistic resources SPUD draws upon, and the remainder of the system works exactly the same as described for REA, as in Section 2, where whole gestures were treated like words. In fact, by the time SPUD searches its lexicon, it will have access to whole gestures; but, instead of requiring a static set of gestures, the gesture planner inserts new lexical entries into the lexicon for each generated gesture.

$$\begin{bmatrix} \text{hand : Left/Right/Both + symmetry symbol} \\ \text{form :} \begin{bmatrix} \text{location :} \begin{bmatrix} \text{pos : location symbols} \\ \text{trajectory : sequence of trajectory elements} \end{bmatrix} \\ \text{extended finger direction : finger direction symbol} \\ \text{palm orientation : direction/rotation symbols} \\ \text{hand shape : handshape symbol} \end{bmatrix} \end{bmatrix}$$

Figure 3. Typed feature structure for gesture.

There are several important differences between SPUD, a sentence planner, or microplanner for language, and the GP, a microplanner for gesture. One key difference is in the nature of the two linguistic resources it draws upon. First, the gesture planner's "lexicon" is not a "gestionary," or library of gestures. Replacing words are instead what we call *form features.* Form features define morphological constituents of gesture, like those observed in Section 3.1. We will derive a set of such form features empirically, through the experiment currently underway, in which we look for patterns in the use and meaning of form features like hand shapes, hand locations, direction of the extended fingers, and orientation of the palm. The semantic formulae associated with these form features are also distinct from the kinds of meanings that words may take on.

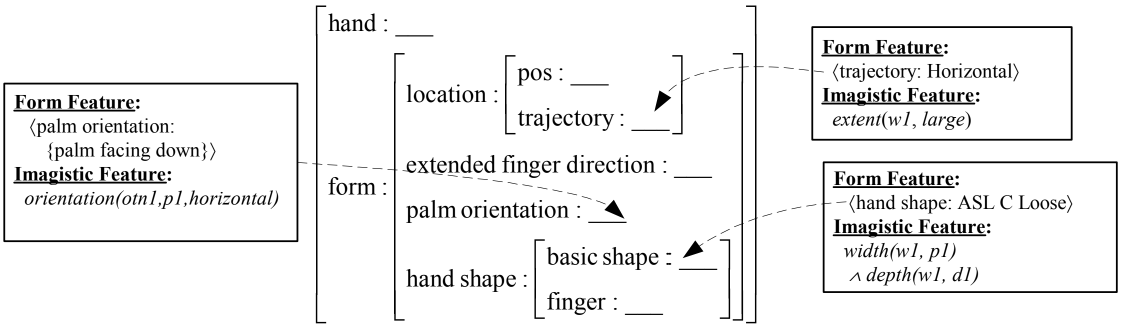

Figure 4. Feature structure filled by gesture form feature entries.

These formulae are restricted to vague, imagistic terms which describe the intrinsic imagistic properties of the form feature, again, like those observed in Section 3.1; e.g. a flat-handshape is generally used for two-dimensional planes, and could thus carry semantic formulae "$width(W, X) \wedge depth(D, Y)$" or "$width(W, X) \wedge height(H, Y)$." Having replaced SPUD's lexical entries with a set of form features, the second part of the linguistic resources to be replaced is the grammar.

While gestures are not hierarchically composed like sentences, they can be described in terms of form features. Therefore, we replace syntactic trees with typed feature structures, as illustrated in Figure 3). Uyechi [27] uses a similar (albeit much more complex) formalism to describe the visual phonology of sign language.

In addition to the process of filling in substitution nodes in a syntactic tree, SPUD's search space is defined by several other heuristics. Similarly, the GP will use heuristics to structure its own search space, delimiting all possible ways to combine a set of form features into a sound form feature structure that defines a realizable gesture. One such heuristic is a set of composition constraints that formalize restrictions over the ways in which different form features combine. Figure 4 illustrates the gesture construction process. Another heuristic, similar to SPUD's preference for keeping sentences as short as possible, is a pragmatic constraint favoring the reuse of feature structure which was successfully used before to express a common set of semantic formulae. This heuristic requires comparison to a record of context, and allows for simulation of McNeill's catchments.

4. CONCLUSION

In this paper, we have proposed a new method for the generation of coordinated language and novel iconic gestures based on a common representation of context and domain knowledge. We apply the SPUD approach to microplanning to gesture planning. Lexical entries are replaced with form feature entries; hierarchical LTAG trees are replaced with feature structures more closely resembling the global and synthetic nature of gesture; LTAG operations are replaced with feature composition rules; and pragmatic constraints are carried over to guide gesture use in context. This model builds on and extends previous work on the REA and MACK systems, both of which have been informed by empirical studies. Continuing this line of research, we are currently collecting further empirical data to refine the theoretical model described in this paper and its application in the NUMACK system, an interactive ECA capable of giving directions in the real-world domain of Northwestern University's campus.

We believe that our approach to microplanning is one step closer to a psychologically realistic model of a central step in utterance formation (cf. [11]). However, while the model presented here comprises two separate but interacting planning processes, a higher degree of interaction may be necessary, as proposed by McNeill [18]. This issue will be addressed by further evaluation of the implementation. Another question to explore is whether a quantitative representation of imagistic information is required. The current approach uses only qualitative representations for several reasons. First, as simplifying assumption, we only employ a single underlying representation as opposed to mixing symbolic and numeric representations. Second, without being linked to context, gestures are vague in meaning. Since we need to reason about the form of gestures before linking them to context, we employ a qualitative representation that facilitates representation and reasoning about such information. Finally, in the empirical analyses conducted so far, a qualitative representation has been adequate to provide the level of description needed to account for the observed behavior. However, we know that motor planning for surface realization of gestural movements in an embodied conversational agent requires a more precise, quantitative specification of the gesture to be performed. Whether this information is also needed in microplanning, is, again, a problem that needs to be illuminated by evaluation of the system.

Acknowledgements

The authors gratefully thank Matthew Stone for helpful comments, and Marc Flury for technical assistance. This work is supported by a fellowship within the Postdoc-Programme of the German Academic Exchange Service (DAAD).

5. REFERENCES

[1] Biederman, I. Recognition-by-Components: A theory of human image understanding. *Psychological Review* 94(a): 115-147, 1987.

[2] Clark, H.H. and Schaefer, E.F. Contributing to discourse. *Cognitive Science*, 13: p. 259-294, 1989.

[3] Clark, H. H. Using Language. Cambridge, UK: Cambridge University Press. 1996.

[4] Cassell, J. and Scott, P.. Distribution of Semantic Features Across Speech and Gesture by Humans and Computers. In *Proc. Workshop on Integration of Gesture in Language and Speech,* 1996.

[5] Cassell, J., Bickmore, T.,Vilhjalmsson, H., Yan, H. More than Just a Pretty Face: Affordances of Embodiment. In *Proc. of IUI 2000*, pp. 52-59. Jan. 4-9, New Orleans, LA, 2000.

[6] Cassell, J., Stone, M. & Yan, H. Coordination and context-dependence in the generation of embodied conversation. In *Proc. of the 11th INLG 2000.* Mitzpe Ramon, Israel, 2000.

[7] Cassell, J., Vilhjalmsson, H. and Bickmore, T. BEAT: the behavior expression animation toolkit. In *Proc. of SIGGRAPH 2001*, pp. 477–486, 2001.

[8] Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Tversky, D., Vaucelle, C., Vilhjlmsson, H. MACK: Media lab Autonomous Conversational

Kiosk. In the *Proc. of Imagina '02* . February 12-15, Monte Carlo, 2002.

[9] Chi, D., Costa, M., Zhao, L. and Badler, N. The emote model for effort and shape. In *Proc. ACM SIGGRAPH*, pp. 173-182, 2000.

[10] Coyne, B. and Sproat, R. WordsEye: An Automatic Text-to-Scene Conversion system. In *Proc. SIGGRAPH 2001*. Los Angeles, CA, 2001.

[11] de Ruiter, J.P. The production of gesture and speech. In McNeill, David (Ed.) *Language and Gesture*. Cambridge, UK: Cambridge University Press, 2000.

[12] Emmorey, K., Tversky, B. & Taylor, H.A., Using space to describe space: Perspective in speech, sign, and gesture. Spatial Cognition and Computation 2:3, pp. 157-180, 2000.

[13] Kopp, S. and Wachsmuth, I. Synthesizing Multimodal Utterances for Conversational Agents. *The Journal Computer Animation and Virtual Worlds*: 15(1), pp. 39-52, 2004.

[14] Landau, B. and Jackendoff, R. What and where in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16, 217-265, 1993.

[15] Levison, L. and Badler, N. How Animated Agents Perform Tasks: Connecting Planning and Manipulation Through Object-Specific Reasoning. Presented at *the AAAI Spring Symposium: Toward Physical Interaction and Manipulation*, March, 1994.

[16] Perlin, K. and Goldberg, A. Improv: A System for Scripting Interactive Actors in Virtual Worlds. In *Proc. SIGGRAPH '96*, New Orleans, LA, ACM Press, 1996.

[17] Marr, D., *Vision*. Freeman, 1982.

[18] McNeill, D. *Hand and Mind: What Gestures Reveal About Thought*. Chicago, IL: Univ. of Chicago Press, 1992.

[19] McNeill, D. Catchments and Contexts: Non-modular factors in speech and gesture production. In McNeill, D. (Ed.) *Language and Gesture*. Cambridge, UK: Cambridge University Press, 2000.

[20] Matheson, C., Massimo P. and Traum, D. Modeling Grounding and Discourse Obligations Using Update Rules, in *Proc. NAACL 2000*.

[21] Nakano, Y., Reinstein, G., Stocky, T., Cassell, J. Towards a Model of Face-to-Face Grounding *Proceedings of ACL 2003.* July 7-12, Sapporo, Japan, 2003.

[22] Reiter, E. and Dale, R. *Building Natural Language Generation Systems*. Cambridge, UK: Cambridge University Press, 2000.

[23] Sowa, T. & Wachsmuth, I. Coverbal Iconic Gestures for Object Descriptions in Virtual Environments: An Empirical Study. In: M. Rector, I. Poggi & N. Trigo (eds.): *Proceedings of the Conference "Gestures. Meaning and Use"*, pp. 365-376, Porto: Edições Fernando Pessoa, 2003.

[24] Stone, M., Doran, C., Webber, B., Bleam, T. and Palmer, M. Microplanning with communicative intentions: the SPUD system. *Computational Intelligence* 19:4, pp. 311-381, 2003.

[25] Stone, M. Intention, interpretation and the computational structure of language. Under review for *Cognitive Science*, 2004.

[26] Traum, D. and Larsson, S. The Information State Approach to Dialogue Management. In *Current and New Directions in Discourse and Dialogue*, ed. J. van Kuppevelt & R. Smith. Kluwer Academic Publishers, 2003.

[27] Uyechi, L.A.N., *The Geometry of Visual Phonology*. Chicago, IL: Univ. of Chicago Press, 1996.

[28] Yan, H. *Paired Speech and Gesture Generation in Embodied Conversational Agents.* Masters Thesis. MIT, Program in Media and Art Sciences, School of Architecture and Planning, 2000.