Towards Integrated Microplanning of Language and Iconic Gesture for Multimodal Output

Stefan Kopp
University of Bielefeld
D-33594 Bielefeld
Germany
+49 521 106 2919
skopp@techfak.uni-bielefeld.de

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

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

ABSTRACT

When talking about spatial domains, humans frequently accompany their explanations with iconic gestures to depict what they are referring to. For example, when giving directions, it is common to see people making gestures that indicate the shape of buildings, or outline a route to be taken by the listener, and these gestures are essential to the understanding of the directions. Based on results from an ongoing study on language and gesture in direction-giving, we propose a framework to analyze such gestural images into semantic units (image description features), and to link these units to morphological features (hand shape, trajectory, etc.). This feature-based framework allows us to generate novel iconic gestures for embodied conversational agents, without drawing on a lexicon of canned gestures. We present an integrated microplanner that derives the form of both coordinated natural language and iconic gesture directly from given communicative goals, and serves as input to the speech and gesture realization engine in our NUMACK project.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing - *Language generation.* H.5.2 [**Information Interfaces and Presentation**]: User interfaces - *Natural language.*

General Terms

Algorithms, Experimentation, Languages, Theory

Keywords

Gesture, language, generation, multimodal output, embodied conversational agents

1. INTRODUCTION

When describing a scene, or otherwise conveying image-rich information, humans make frequent use of *iconic gestures,* movements of the hands and arms that express spatial, image-evoking information and resemble the object or event being referred to. For example, when somebody is asked how to find a building in a city, it is common to see the direction-giver depicting significant landmarks with the hands—the fork where one road joins another, the shape of remarkable buildings, or their spatial relationship to one another. Figure 1 shows an example from an ongoing study on spontaneous gesture in direction giving. Here, the speaker has just said "*If you were to go south*", then, while making the shown gesture, he says "*there's a church*". The gesture imparts visual information to the description, the shape of the church (left hand) and its location relative to the curve of a road (represented by the right arm), and this meaning is instrumental to the understanding of the scene. Computer systems that support human-like communication in similar domains cannot afford to ignore the benefits of such coordinated language and gesture use. However, since gesture is spontaneously produced by speakers, and since gestures do not derive from a lexicon the way words do, multimodal systems research still struggles with the autonomous generation of paired gestural and verbal output in embodied agents. At the same time, a believable interface agent must be able to conceive, represent, and convey a vast number of meanings that multimodal interaction can express (cf. [22]). An important task, then, in creating conversational interfaces that really communicate multimodally, is to discover the rules that link the form of communicative signals to meanings. This includes rules that state how suitable, new signals can be created on the fly.

In this paper, we present a new approach to generating multimodal utterances, where the form of paired natural language and iconic gesture is derived directly from given communicative goals, i.e., semantic facts to be conveyed. The hallmark of our approach is a framework to analyze the information presented by iconic gesture into semantic units, *image description features*, and to link these features to discrete form features (hand shape, trajectory, etc.) of gesture. This research is carried out in the ongoing NUMACK project, an embodied conversational agent that answers questions about locations and buildings on Northwestern University campus, and provides directions to each. After discussing related work in the following section, Section 3 introduces our framework, describes its underlying theoretical assumptions, and discusses its empirical basis. Based on this framework, and by extending existent models of natural language generation, Section 4 presents an integrated microplanner for constructing the surface structure of language as well as the form of iconic gesture for NUMACK's utterances on the fly. We rely on an ongoing study on direction-giving dialogues to refine this model and to inform the underlying resources to be integrated within this framework.

ICMI'04, October 13–15, 2004, State College, Pennsylvania, USA.

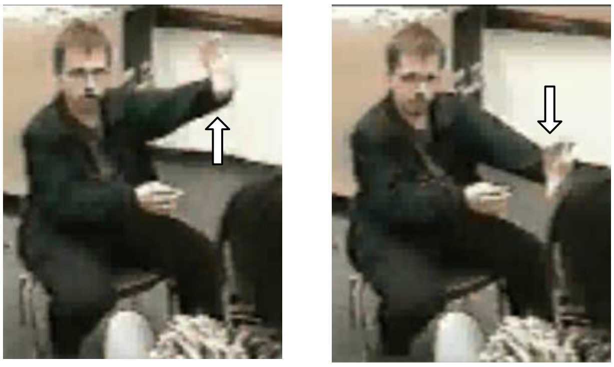

Figure 1. Coverbal gesture on "*There's a church*."

2. RELATED WORK

Many researchers have considered human gesture as a modality of communication. In computational research, for the most part, researchers have abided by pattern classification paradigms in recognizing and interpreting it. The first system to tackle the problem of analyzing a gesture's morphology into its image content, and to exploit the gesture's iconicity in order to determine the object the user is referring to in the context of speech, was the *ICONIC* system [14]. Gesture form was extracted into separate features, and basic hand postures were compared with object shape components like corners, flat sides, major and minor axes or default directions. This approach was extended by Sowa & Wachsmuth [25] who employ a quantitative representation (imagistic description trees) that abstracts both the shape of an real-world object as well as the virtual shape depicted by an iconic gesture, or a sequence thereof, into a spatial arrangement of significant axes. The gesture's referent is then determined by comparing these imagistic representations.

Much existing work has addressed the automatic generation of coordinated language and visualization for complex spatial information (e.g. [27][13][10]). These systems tackle the related problem of distributing semantic content across different output modalities and of coordinating them correctly. Research on multimodal generation of language and gesture has been carried out primarily in the context of embodied conversational agents (cf. Pelachaud & Poggi [22]). Traum & Rickel [28] present a model of dialogue acts for spoken conversation that incorporates non-verbal behavior into its representation as well as accounts for a representation of the discourse state of these dialogue acts. This work is related in that it deals with discourse state of non-verbal behavior [24], but it does not consider questions of generating these behaviors. Nijholt et al. [20] discuss architectural issues for multimodal microplanning and the factors influencing modality choice, but adhere in their proposed model to selecting iconic and deictic gestures from a lexicon; the issues of iconicity of gesture and their underlying semantics are not considered. To date, the *REA* system [3] represents the most elaborated work on the automatic generation of natural language and gesture in embodied conversational agents (ECAs). Using the SPUD system [26] for planning natural language utterances, *REA* was able to successfully generate context-appropriate language and gesture, relying upon empirical evidence [1][28] that communicative content can be defined in terms of semantic components, and that different combinations of verbal and gestural elements represent different distributions of these components across the modalities. This approach was able to account for the fact that iconic gestures are not independent of speech but vary with the linguistic expression they accompany and the context in which they are produced, being sometimes redundant and sometimes complementary to the information conveyed in words. However, whole gestures were treated exactly like words, associated to syntactic trees by a specific grammatical construction, the SYNC structure, and gesture planning only extended as far as the selection of a complete gesture from a library and its context-dependent coordination with speech. This does not allow for the expression of new content in gestures, as is possible in language with a generative grammar. Gao [9] extended the *REA* system to derive iconic gestures directly from a 3D graphics scene. He augmented the VRML scene description with information about 3D locations of objects and their basic shapes (boxes, cylinders, spheres, user-defined polygons, or composites of these), which were mapped onto a set of hand shapes and spatial hand configurations. This method allows for deriving a range of new gesture forms, but it does not provide a unified way of representing and processing the knowledge underlying coordinated language and gesture use.

The fact that previous systems usually draw upon a "gestionary", a lexicon of self-contained gestures, is also a consequence of the use of canned gesture animations. Although previous systems, e.g. *BEAT* [4], were able create nonverbal as well as paraverbal behaviors—eyebrow raises, eye gaze, head nods, gestures, and intonation contours—and to schedule those behaviors with respect to synthesized text output, the level of animation was always restricted to predefined animations. Sometimes, motor primitives were used that allowed for some open parameters (e.g., in the *STEVE* system [24] or *REA* [3]), were adjustable by means of procedural animation (*EMOTE* [5]), or could be combined to form more complex movements (e.g. [21]). For example, [15] presented a generation model that assembles gestural motor behaviors on the fly, entirely based on specifications of their desired overt form. This method allows for greater flexibility with respect to the producible forms of gesture, which is clearly a prerequisite for the level of gesture generation targeted here.

3. A FEATURE-BASED FRAMEWORK FOR ICONIC GESTURE

Iconic gesture depicts visual information about an object or action being referred to, and its communicative functions have been reported to range from conveying the extent of a building to specifying the direction or manner of an action, or the viewpoint from which it is described [18][1]. Whatever it conveys, iconic gestures are never independent of speech, for their features are found not only to map onto the visual component of a scene or event, but also to depend on the linguistic context in which they are produced. Further, gestures can be found to be either—at least partially—redundant with speech or to contribute information that is completely complementary to what is encoded in speech [3][1]. Research has indicated that, in any case, listeners do attend to the information conveyed in gesture, even when it contradicts the information conveyed by speech [2].

Since iconic gestures communicate in virtue of their resemblance with the information they depict, there are fundamental differences between the "semantics" of gesture and the lexical semantics posited for words. Words are arbitrarily linked to the concepts they represent, gestures are not. While a word may have a limited number of possible meanings, an iconic gesture without context is *vague* from the point of view of the observer, i.e., it displays an image that has a potentially countless number of interpretations in isolation. For example, the gesture shown in Figure 1 can be used to depict anything from the movement of an elevator to the shape of a tower. That is, it does not make sense to say that a gesture—observed as a stand-alone element separate from the language it occurs with—has semantics in the same way as gesture does when interpreted within linguistic context. We therefore introduce a qualitative, intermediate level of analysis to describe the inherent content (semantics) of an iconic gesture, without context. And we propose that the gesture's semantics on this level should describe meaningful features of shape, spatial properties, and spatial relationships. We call these descriptors *image description features* (henceforth, IDFs).

Our approach builds on the assumption that one can trace communicative content from this mediating level to its surface realization in gesture morphology. If iconic gestures are indeed communicative, people must be able to recover and interpret their meaning, and a reliable system for this requires some systematicity in the way gestures are used. We hypothesize that this system comprises prevalent patterns in the ways the hands and arms are used to create iconic, gestural images of salient, visual aspects of objects/events. We further assume that pervasive patterns can be found as to how distinct IDFs are associated with primitive *form features* of gesture morphology, i.e., hand shapes, orientations, locations, and movements in space, or combinations thereof—and we believe that such patterns can be used to characterize the ways human speakers derive novel iconic gestures for objects they may be describing for the first time. For example, in the utterance data shown in Figure 1, the subject's right hand is held in place from the previous utterance and represents the curve in a road, anchoring the frame of reference. In the left-hand gesture, we find three form features: the flat hand shape with slightly bent fingers (similar to the American Sign Language (ASL) sign *B spread*), the vertical linear trajectory, and the hand location relative to the right-hand gesture. Each form feature corresponds to one or more IDFs in virtue of the resemblance of the former to the latter. The relatively flat handshape resembles a flat shape; or in more descriptive spatial terms, a two-dimensional, planar shape in a certain orientation. The vertical, linear trajectory shape corresponds to a feature that marks a vertical extent. Finally, the gesture location corresponds to a spatial location in relation to the frame of reference.[1] All three IDFs in combination define an upright plane with a significant vertical extent, in a particular orientation and location.

By following such patterns between IDFs and form features, we can plan iconic gestures from an intended interpretation, which in turn must take account of the context of simultaneous speech. Only when an iconic gesture is placed in linguistic context, does the set of possible interpretations of the IDFs become so constrained as to make it unique. In the example in Figure 1, we infer from the indefinite noun phrase "*a church*" that the IDFs represent spatial information about *the referent* of the expression, namely *a church*. Linking the underspecified, imagistic features to this specific referent makes a successful interpretation possible. The depicted upright plane becomes the wall of the church, viewed relative to the location of the road, and the vertical trajectory emphasizes the salient, vertical dimension, now corresponding to the *height* of the wall. Overall, we infer that the communicative intention of the subject was to introduce a church, which has a tall, upright wall, and which is located near the curve of the road.

3.1 Empirical Basis

Our hypothesis originates in empirical evidence from previous literature found in several domains, suggesting patterns in the form and function of iconic gestures with respect to expressing spatial information and communicating meaning more generally. For example, Sowa & Wachsmuth [25] report that one can find consistencies in the ways the fingers are used to trace a shape and that both palms may be held facing each other to illustrate an object's extent. Unlike language, in gesture multiple form features may be combined to express multiple spatial aspects (e.g., extent and shape) simultaneously. Emmorey et al. [8] observed that depictions of complex spatial structures are broken down into features that are then built up again by successive gestures. The fact that a single spatial structure is referred to across gestures (for example, a winding road) is signaled by spatial coherence; that is, the gestures employ the same viewpoint, size scale, and frame of reference, as indicated by a constancy of hand shape, trajectory and position in space. Sometimes, the frame of reference (e.g. relative to the winding road) is explicitly anchored in gesture space by one hand, and then held throughout while the other hand describes additional landmarks at appropriate relative locations. McNeill & Levy [17] found positive and negative correlations for the association of distinct "kinesic" features in gesture, like fingers curled, palm down, or motion upwards, with semantic features of the motion verbs the gestures co-occurred with. For example, verbs with a horizontal meaning feature tended to co-occur with gestures with a sideways movement, but almost never with downward motion.

To test our hypothesis, we have collected and are currently analyzing videotapes of 28 dyads (more than five hours of dialogue) engaging in direction-giving. We predict that the data will reveal patterns, i.e. sets of one or more form features being used to convey sets of one or more IDFs; i.e., IDF-form feature mappings. If we can show that similar gestural form is generally used to depict visually similar things, we will have evidence supporting our hypothesis.

In a preliminary analysis, we selected sample dyads with directions given to different locations, and then selected all gestures with a particular form feature, or combinations thereof, *regardless* of what the remaining form features looked like. For these gestures, we then looked at correlations between form and first-level meaning (IDFs). When analyzing the pairings of form features and IDFs, we found that 67% of the gestures with a linear trajectory (N=48) referred to objects, and that the gestures tend to depict a significant axis with the linear movement (e.g., run of a street, extent of a field, transverse overpass). Likewise, 80% of the gestures with a flat hand shape and the palm oriented vertically (N=45) referred to objects whose shape comprises an upright plane (walls, stop sign, window, etc.). Finally, we looked at combinations of a flat hand shape with the palm oriented in a way such that the thumb points upwards (N=61). Such gestures occur with a high frequency in our corpus, and they seem to be the predominant way to illustrate change in location or/and

[1] In the spatial language literature (cf. [16]), a distinction is drawn between a located object or *figure* and a reference object or *ground*.

orientation. Although the two form features combine with various movements and locations in space, still 85% of the gestures referred to directed actions (go, keep on, take a, look, etc.), and the gesture always ended up with fingers pointing directly in a direction or to a location the action is directed at. Overall, these results suggest a significant correspondence between combinations of form features and IDFs—the patterns we were looking for. In addition, it appears that in gestures of the last category the trajectory of the hands depicts the concrete path or direction of the action described—a fairly unmediated and direct iconic mapping. That is, these gestures fuse iconic with deictic aspects, and hand shape and palm orientation, on the other hand, appear to be more "conventionalized" in their indexing of locations and directions (cf. [7]). Note, again, that these results are preliminary and must be scrutinized and generalized in further analyses. Our hypothesis will only be verified if, in addition to the examination of larger data samples resulting in statistical significance, the opposite direction is also evaluated, that is, positive evidence is found that objects with similar IDFs are depicted by gestures with similar form feature.

4. GENERATING MULTIMODAL UTTERANCES

Based on the framework described in the previous section, linking form features to IDFs, we can approach the encoding of a given communicative intention into gesture morphology. We extend a Natural Language Generation (NLG) model to generation of natural language and iconic gesture (henceforth, NLGG). Commonly, NLG architectures are modular, pipeline architectures, broken down into three subtasks—content planning (also known as text or document planning), microplanning and surface realization (in that order) [23]. In ordinary language, the work done by these three subsystems boils down to, respectively, figuring out what to say, figuring out how to say it, and finally, saying it. Here we focus on microplanning, the second stage of the NLGG pipeline, where domain knowledge must be recoded into linguistic and gesture form. But, since all these stages are crucially linked we will outline some prerequisites to be met by the other stages.

4.1 Modeling Content

At the level of content planning, our NLGG model requires a rich representation of domain knowledge that pays attention to the affordances of both language and gesture as output media. In our present project on direction giving, most of this content is spatial information about actions, locations, orientations, and shapes of landmarks. We can incorporate findings from literature on spatial language (e.g., [16][11]) to understand its representational abilities, and to allow for accurate generation of such language. Unfortunately, there is no such literature offering a precise description of the gestures accompanying spatial language in this domain, or of the nature of the interaction between the two modes. Therefore, we are working to inform our representation with findings from our own current study, as described in Section 3.

To model natural language, we require two theoretical levels of abstraction, with corresponding layers of formal (or symbolic) representations. For example, for a NLG system to refer to an object as "tall", there are two theoretically distinct levels: First, the concept or property of *tallness* can be formalized as a simple logical formula like *tall*(*X*), where *tall* is a predicate symbol representing the concept, and *X* is an open variable (capitalized in Prolog notation) which can be instantiated to another, ground symbol, representing a particular discourse referent (e.g., *tall*(*church*) or *tall*(*john*)). Second, this formula must be associated with the string "tall" representing the word itself. For language this mapping between words and meaning is sufficient, but for gesture which depicts a concept by virtue of resemblance, this level of granularity is too coarse. Abstractly, *tallness* requires a more fine-grained specification, in terms of the intrinsically three-dimensional, spatial nature of this property. Spatial properties could describe tallness as holding of an object when the extent of its vertical axis is longer than its other axes, or more likely it is long relative the vertical axes of some other relevant objects (e.g., a man might be tall relative to some other men standing nearby), or relative to some stereotype. Thus we use the level of IDFs, an intermediate level of abstraction, to represent spatial properties that can be displayed by gesture. If the concept of tallness is represented as *tall*(*X*), and its spatial description is represented as a set of IDFs, we can then map these IDFs onto form features, and this iconic gesture can be used to refer to the concept.

This example motivates IDFs and conceptual/semantic knowledge as two different kinds of knowledge, emphasizing the different levels of abstraction needed to exploit the representational capabilities of the two modalities, and meriting separation into two ontologically distinct levels. However, at the same time, we posit that both language and gesture should utilize one common representation of content and context, working together to express information as parts of one communicative system [18]. Our approach thus maintains a *single*, *common* representation system, in terms of qualitative, logical formulae, encompassing both kinds of domain knowledge needed. We base this system on a formal hierarchical ontology, including objects (e.g. buildings, signs), regions (e.g. parking lots, the north side), shapes, locations and directions (both relative and absolute), qualitative extents (long, short, large, small), events (go, turn), etc., all connected using taxonomic, partonomic, and spatial relations (isa, part of, in, on, etc.). We have begun building such a detailed ontology for parts of Northwestern University campus, as it will ultimately be required for the full working ECA system. Content plans are then specified in terms of these entities and relations.

We focus here on microplanning, so discussion of how domain knowledge would be selected and structured into coherent directions is beyond the scope of this paper. But, microplanning follows naturally and expeditiously from the design of the knowledge, and we give here an example of a content plan our system is able to process. Formalized in logic, it comprises both kinds of knowledge required for employing language and iconic gesture in instructing someone that she will see a particularly shaped building ("Cook Hall") on her right:

instruction(*e*2). *see*(*e*2,*user*,*cook*,*future*,*place*(*on*,*right*)).
tense(*e*2,*future*).
name(*cook*,*cook_hall*). *type*(*cook*,*building*).
place(*on*,*right*). *rel_loc*(*cook*,*user*,*right*). *shape*(*dim*,*vert*,*cook*).
shape(*primary_dim*(*longit*,*cook*)).
shape(*dim*,*longit*,*cook*).

Figure 2: Sample content plan.

From here, we move on to microplanning, and we will see how our model constructs an utterance that conveys all communicative goals comprised by this content plan.

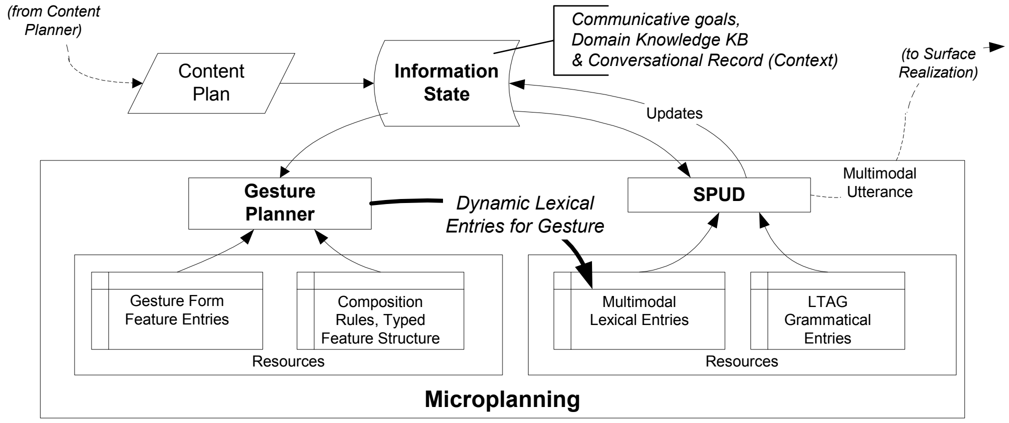

Figure 3. Multimodal microplanning architecture.

4.2 Multimodal Microplanning

As mentioned, a multimodal microplanner must link domain-specific representations of meaning to linguistic form and gesture form. As we have just shown, each modality is significantly different, requiring different kinds of information, providing different representational capacities, and conveying information in different ways. Therefore, NLGG requires different models of how each modality is able to encode content. We thus propose the addition of a new subsystem for *gesture planning* within the microplanning stage of NLGG, as illustrated in Figure 3. This new component, the gesture planner (GP), is responsible for planning the form of new gestures to encode a set of one or more input IDFs. That is, the GP is itself a microplanner, addressing the problem of recoding content into form, but this time on a feature level, from IDFs to morphological form features. We employ SPUD[2], a grammar-based natural language generation micro-planner [26], to connect content to linguistic forms, selecting words and syntactic structures from a set of linguistic resources, and composing utterances to express communicative goals within the constraints of context. In previous work [3], SPUD's linguistic resources were extended to include a set of predefined gestures, from which it drew upon to express its communicative goals. We follow this same strategy here, using SPUD to compose full, multimodal utterances via a single, uniform algorithm. But, instead of drawing upon a static set of predefined gestures, we add the GP into the pipeline: before calling SPUD, the GP plans gesture(s) to express IDFs. These dynamically planned gestures are then incorporated into SPUD's linguistic resources (now *multimodal* resources) and utilized in the same way as in [3]. In the remainder of this section, we summarize the SPUD approach to NLG, before showing how it is extended to planning gesture in the next section.

Microplanning is often dealt with as three distinct tasks or stages: *lexical choice*, *aggregation,* and *referring expression generation* [23]. In the SPUD system, Stone et al. [26] advocate a uniform approach to microplanning, framing it as one search task wherein utterances are constructed iteratively, from an input specification of a set of linguistic resources (grammar) and a knowledge base (KB). Combined with a set of facts to be communicated by a generated utterance, or communicative effects, the input specifications define the search space for the task. Each state in the search space comprises a stage in the construction of a potential utterance. The grammar includes a set of syntactic constructions and a set of *lexical entries*. Syntactic constructions are data structures, each containing a syntactic structure (a tree, formalized using Lexicalized Tree Adjoining Grammar, or LTAG [12]) and an optional pragmatic condition expressing constraints on use of the construction relative to the discourse context. Lexical entries are data structures containing: a lexical item (word), a set of logical formulae defining meaning (semantics) and conditions for use in conversation (pragmatics), a list of predicate-arguments for the formulae (parameters), and a list of trees that the word can "anchor". The KB consists of facts about the domain and conversational context. All facts are explicitly labeled with information about their conversational status, e.g. whether the fact is *private* or *shared*, constraining decisions about what information the system must assert as new to the user, and what it can presuppose as information in the common ground [6]. Therefore, communicative effects, the messages to be conveyed from the content plan, are usually labeled private.

Solution of the three microplanning problems is never dealt with explicitly, but occurs naturally during the course of the SPUD generation process. Working towards a complete, grammatical structure, each stage in an LTAG derivation (and in the SPUD search algorithm), involves the selection of a new syntactic structure to add. Since trees are lexicalized, every syntactic choice is tied to a choice of a word. This is why each lexical entry includes a list of trees it can anchor, and it allows for simultaneous construction of syntactic and semantic structure, in a single, tightly integrated process. Decisions in lexical choice are dealt with primarily in terms of a heuristic cost function, as part of the calculation of a state's overall cost in the greedy search algorithm. Words are chosen such that their semantics allows for the assertion of the maximum number of communicative effects to be achieved per state. The result is an economical, Gricean approach to generation, and the solution of the aggregation problem, as the content is maximized while utterance length is minimized. When one word cannot identify a referent uniquely, additional words (e.g. modifiers) are iteratively added until the referent can be picked out from potential distractors. In this way, SPUD generates referring expressions [23]. Additionally, for each state, the system maintains a representation of the utterance's intended interpretation, or *communicative intent*, a record of inferential links made in connecting the semantic and pragmatics associated with linguistic terms, to facts about referents in the world, as recorded in the KB.

To make this explanation more concrete, Figure 4 shows a snapshot in the derivation process of an utterance that encodes the

[2] SPUD stands for *Sentence Planning Using Descriptions*.

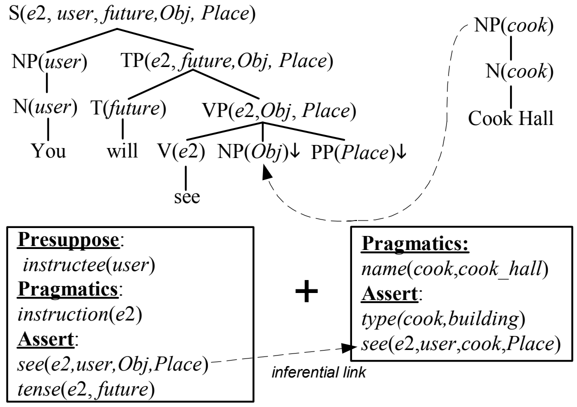

Figure 4. Snapshot of SPUD utterance construction

messages of the content plan in Figure 2. The system works toward assertion of the formula *see*(*e*2,*user,cook,place*(*on,right*)), a communicative effect. Each node in the tree is associated with a set of *predicate arguments*, a formal convention which allows the system to keep track of the links between the constituents of the syntactic structure, and the arguments of its corresponding logical formulae. At this stage in the derivation, the tree for a transitive verb has already been selected, as well as its anchor *see*, with corresponding semantics *see*(*Event,Agent,Obj,Place*), denoting an *Event* where *Agent* sees *Obj* at *Place*. As the system works towards assertion of the aforementioned communicative effect, each of these open variables is instantiated to the specified ground terms, or discourse referents, seen in the formula.

Here, an NP tree is chosen to complete the next stage in the LTAG derivation process. This selection fills in the syntactic object argument for the transitive verb *see*. The new syntactic structure must be anchored by a noun phrase whose semantic predicate argument matches the *Obj* argument of the communicative effect, namely *cook*. Searching through all possible lexical entries that can anchor an NP tree will produce a list of nouns, carrying a semantic formula that defines an object of a particular type, e.g. *type*(*X*,*building*) or *type*(*X,sign*). The choice here is guided and constrained by the other input communicative effects for this utterance, seen in the content plan (Figure 2). Since the only *type* formula the system must assert is *type*(*cook,building*), selection of the lexical entry for "Cook Hall" is the best choice at this stage. So, the new tree is inserted at the NP node in the tree, and the variable *Obj* is bound to the term *cook*, an inferential link from the semantic formula associated with "Cook Hall" to a particular discourse referent, representing the object in the world. In our current project, we use a fast, lightweight, Prolog implementation of SPUD, wherein inference from open variable parameters to particular referents in the KB is achieved via Prolog unification.

The SPUD system provides an elegant solution to the microplanning problem, solving what are often dealt with as a set of artificially demarcated subproblems in a single, uniform approach. The primary drawback of the system is its dependency upon rich linguistic resources and representations of domain knowledge, which causes a knowledge acquisition bottleneck. However, it is precisely this tight integration between the representation system used for domain knowledge and natural language semantics, and the flexibility afforded by a logic-based representation language that makes our approach possible.

4.3 Gesture Planning and Integration

The Gesture Planner system is inspired by the SPUD approach, drawing upon a bipartite input specification of domain knowledge, plus a set of entries to encode the connection between semantic content and form. Using such data structures, we are able to achieve the same kind of close coupling between gesture form and meaning, allowing for efficient, incremental construction of gestures and maintenance of inferential links from abstract meaning (logical) formulae to specific discourse referents. For the GP, we formalize the form-meaning coupling in a set of *form feature entries*, data structures that connect IDFs to morphological form features. These form feature entries implement the patterns that we find in our empirical data, and they may contain "clusters" of features on either side, i.e., conjunctions of IDFs as well as combinations of morphological form features. Also, we use these entries to encode the ways in which the function of a gesture (e.g., deictic) influences its form (e.g. the hand shape) through conventionalized patterns, again, as suggested by our empirical data.

The GP searches for all combinations of form feature entries that can realize an input set of IDFs, the desired communicative effects. Since iconic gesture is not governed by a hierarchical system of well-formedness like language, a formal grammar is inappropriate for this purpose (cf. [18]). Instead, we employ an algorithm similar to feature structure unification to combine form features, whereby any two form features may combine provided that the derived feature structure contains only one of any feature type at a time. Through iterative application of this operation, the GP builds up gestures incrementally until all the desired communicative effects are encoded. Figure 5 shows a state in the generation of a gesture, composed to depict the IDFs from the content plan in Figure 2. Location and hand shape have already been inserted, the latter according to the pattern we have observed in our data, namely, the use of a flat hand shape (ASL sign *5*) and a vertically oriented palm for depicting a wall (defined as a planar, upright surface in our ontology, which also tells us that *cook*, being a building, has walls). The same pattern now informs the palm orientation, together with the location of the object (*cook*) to be depicted.

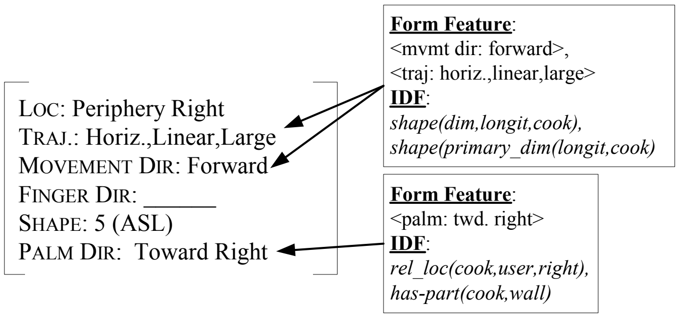

Figure 5: Example of form features entries filling a gesture feature structure.

Similar to SPUD's pragmatic constraints on the way language is used in context, the GP process can be guided by composition constraints on all possible ways to combine a set of form features into a feature structure that defines a realizable gesture. Such composition constraints could formalize restrictions over the ways in which different form features combine, and could, for example, be utilized to favor the reuse of feature structures that have been successfully used before to express a common set of semantic formulae. This would require comparison to the KB's record of context, and allows for simulation of what McNeill has called catchments [19], the maintenance of a certain form feature in gesture to indicate cohesion with what went before.

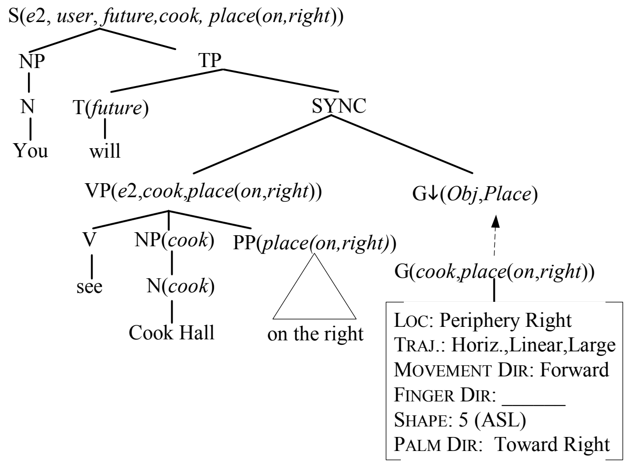

Figure 6: Insertion of the gesture into the utterance tree.

In our current Prolog implementation, the GP simply returns all possible form feature combinations, i.e., it delivers all gestures that could take on communicative work by encoding some or all of the desired communicative effects. It is up to SPUD to choose the gesture that, when combined with appropriate language, allows for the most complete intended interpretation in context (cf. Section 4.2). To this end, all the dynamically planned gestures are incorporated into SPUD's linguistic resources and utilized the same way as in [3]. The SYNC construction pairs a syntactic constituent and a gesture feature structure under the condition that their predicate arguments are connected to the same discourse referents, achieving coordination of meaning in context. In addition, it poses a constraint of temporal surface synchrony between both elements. Possible ways in which language and gesture can combine are then represented by a set of such SYNC constructions in the GP's resources. The SPUD algorithm chooses constructions and the gesture form feature structures in the same way as explained in Section 4.2. Figure 6 shows how the gesture, derived for the IDFs in the content plan from Figure 2, is incorporated into a multimodal utterance, analogous to Figure 4.

Finally, the description of the resultant multimodal utterance is converted into an XML tree, and then passed on to the next and final stage of our NLGG pipeline, surface realization. Note that the GP may output an underspecified gesture if a form feature is not a meaningful aspect of a gesture. These features remain open during gesture planning, and will be default to anatomically suitable values during surface realization.

4.4 Surface Realization

For an embodied conversational agent, surface realization must turn the results of microplanning into morphologically and phonologically-specified synthesized speech and intonation, expressive gesture animations for a graphical avatar body, and schedule both into synchronized multimodal output. With respect to gesture generation, the microplanner must be able to come up with form specifications of potentially novel iconic gestures. That is, we cannot rely on canned gesture animations, but need a module for calculating appropriate animations on the fly. These problems have been tackled in previous generation engines (see Section 2), and we combine modules from the *BEAT* system [4] as well as the *MAX* system [15] to form a realization pipeline. Our pipeline starts out with a XML specification of the utterance, as outputted by the microplanner. The rule-based *BEAT* models for suggesting and filtering conversational behaviors are applied, first, to augment the description with apposite nonverbal and paraverbal behaviors (intonation, eyebrow raise, head nod, posture shift, etc.). Then, employing the *Festival* system for text-to-speech conversion, the *MAX* system is used to schedule all verbal and nonverbal behaviors, and to finally create all the animations required to drive NUMACK's body and face as specified. For the hand gestures, this includes a motor planning stages that processes the form definition originating from micro-planning and timing constraints set up during scheduling, and turns them into applicable motor programs (see [15] for details).

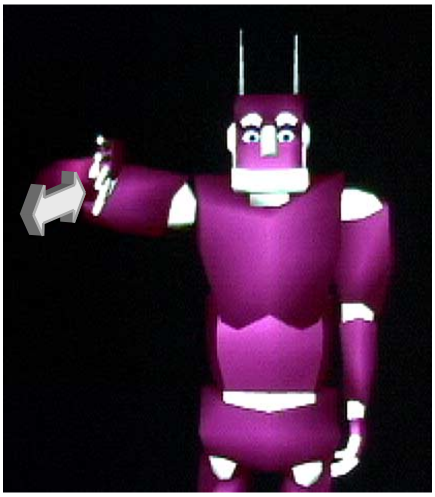

Figure 7: NUMACK realizing the utterance in Figure 6; "*You will see Cook Hall on your right*".

5. CONCLUSION

In this paper, we have proposed an approach to the automatic generation of language and iconic gestures for multimodal output in the NUMACK system, an interactive ECA capable of giving directions in the real-world domain of Northwestern University campus. Building on the results from an ongoing study on gesture and speech in direction-giving, we have presented a framework to link image description features to discrete morphological features in gesture. We have applied this framework to an integrated, on-the-fly microplanning process that derives coordinated surface forms for both modalities from a common representation of context and domain knowledge. In extending the SPUD micro-planning approach to gesture planning, lexical entries were replaced with form feature entries; LTAG trees were replaced with feature structures, more closely resembling the global and synthetic nature of gesture; and pragmatic constraints were carried over to guide gesture use in context. In our current implementation, NUMACK is able to produce, in real-time, a limited range of directions, using semantically coordinated language and gesture. Figure 7 demonstrates the utterances generated from the content plan in Figure 2. We are continuing to analyze our empirical data to refine our model, and to find further patterns in the way iconic gesture expresses visual domain knowledge in order to extend the system's generation capabilities. We believe that our approach to microplanning is one step closer to a psychologically realistic model of a central step in utterance formation. However, a range of open questions still need to be investigated, and evaluation of our system will help us shed light on these problems. Such questions are as to whether a higher degree of interaction might be necessary between the two separate, but interacting planning processes for language and gesture; or whether one unified, qualitative, logic-based representation is sufficient to represent the knowledge required for planning the surface structure of both modalities, while we know that final motor planning of the avatar movement requires a precise, numerical specification of the gesture to be performed.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

[1] Cassell, J. & Prevost, S. Distribution of Semantic Features Across Speech and Gesture by Humans and Computers. In *Proc. Workshop on Integration of Gesture in Language and Speech*, 1996, Wilmington, DE.

[2] Cassell, J., McNeill, D. & McCullough, K.E. Speech-Gesture Mismatches: Evidence for One Underlying Representation of Linguistic and Non-Linguistic Information. *Pragmatics and Cognition* 7(1): 1-33, 1999.

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

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

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

[6] Clark, H. H. Using Language. Cambridge Univ. Press, 1996.

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

[8] 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.

[9] Gao, Y. Automatic extraction of spatial location for gesture generation, Master thesis, MIT Dept. of Electrical Engineering and Computer Science, 2002.

[10] Green, N., G. Carenini, et al. A Media-Independent Content Language for Integrated Text and Graphics Generation. *Workshop on Content Visualization and Intermedia Representations at COLING/ACL '98*, Montreal, 1998.

[11] Herskovits, A. *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English*. Cambridge University Press, 1986.

[12] Joshi, A.K. An Introduction to Tree Adjoining Grammars, In A. Manaster-Ramer (ed), *Mathematics of Language,* Amsterdam: John Benjamins, pp. 87–114, 1987.

[13] Kerpedjiev, S., Carenini, G., Roth, S. & Moore, J. D. AutoBrief: A multimedia presentation system for assisting data analysis. In *Computer Standards and Interfaces,* 18: 583-593, 1997.

[14] Koons, D.B., Sparrell, C.J., & Thorisson, K. Integrating simultaneous input from speech, gaze, and hand gestures. In M. Maybury (ed), *Intelligent Multimedia Interfaces*, pp. 252-276. Menlo Park, CA: MIT Press, 1993.

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

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

[17] McNeill, D. & Levy, E. Conceptual representations in language activity and gesture. In R. Jarvella, & W. Klein (eds.): *Speech, Place, and Action*, John Wiley & Sons, 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 D. McNeill (ed.): *Language and Gesture*. Cambridge University Press, 2000.

[20] Nijholt, A., Theune , M. & Heylen, D. Embodied Language Generation, In O. Stock & M. Zancanaro (eds): *Intelligent Information Presentation*, Kluwer, 2004.

[21] Perlin, K. & Goldberg, A. Improv: A System for Scripting Interactive Actors in Virtual Worlds. In *Proc. SIGGRAPH '96*, pp. 205-216.

[22] Pelachaud, C. & Poggi, I. Multimodal Embodied Agents, In Autonomous Agents Workshop *Multimodal Communication and Context in Embodied Agents*, pp. 95-99, 2001.

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

[24] Rickel, J., Marsella, S., Gratch, J., Hill, R., Traum, D. & Swartout, W. Toward a New Generation of Virtual Humans for Interactive Experiences. *IEEE Intelligent Systems* 17(4): 32-38, 2002.

[25] Sowa, T. & Wachsmuth, I. Coverbal Iconic Gestures for Object Descriptions in Virtual Environments: An Empirical Study. In M. Rector, I. Poggi & N. Trigo (eds.): *Proc. "Gestures. Meaning and Use"*, pp. 365-376, 2003.

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

[27] Towns, S., Callaway, C. & Lester, J. Generating Coordinated Natural Language and 3D Animations for Complex Spatial Explanations. In *Proc. AAAI-98*, pp. 112-119, 1998.

[28] Traum, D. & Rickel, J. Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds. In *Proc. Autonomous Agents and Multi-Agent Systems*, pp. 766-773, 2002.

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