18

Implementing a non-modular theory of language production in an embodied conversational agent

Timo Sowa, Stefan Kopp, Susan Duncan, David McNeill, and Ipke Wachsmuth

18.1 Introduction

Producing language in spoken discourse is virtually impossible without gestures. Growth Point (GP) theory (McNeill 1992, 2005; McNeill and Duncan 2000) articulates a cognitive model of language production that acknowledges the crucial role of embodiment for speaking in that gestures and speech both are considered integral to language. The model is founded on empirical examination of extended natural discourse, emphasizing fine-grained analysis of synchronous, coexpressive speech and gestures.

One, increasingly popular, method to test and to refine cognitive models of language production are computer simulations of multimodal behavior that figure in embodied conversational agents, hereafter ECAs (Cassell *et al.* 2000; see also Poggi and Pelachaud, this volume). Since an ECA always "embodies" a theory, varying the technical model according to different theoretical assumptions has direct impact on its communicative behavior. The effects of manipulating model parameters may then be compared to observations of human behavior and can further inform the modeling effort. On the other hand, confronting an ECA with theoretical psychological concepts like those implied by GP theory can elucidate limits on the computational modeling of human functioning, and can motivate further improvements of ECAs and their communicative behavior.

The aim of this chapter is to discuss and assess the feasibility of operationalizing GP theory's model of language production in an ECA. GP theory and computational ECA models have so far been considered to be largely contradictory in a number of central assumptions, the most crucial being the rejection or adoption of a modular structure of the language production system. We first sketch the cornerstones of non-modular GP theory and its empirical basis. Second, we overview the gesture and speech production models that are currently realized in ECAs, and we discuss their potential and limitations with respect to which characteristics of natural speech and gesture they can account for. Such agent architectures are largely inspired by modularist views of speech production, such as Levelt's "Blueprint for the Speaker" (Levelt 1989). We contrast these theoretical

assumptions with the assumptions and implications of GP theory, focusing on the model architectures of the communicative agents *Max* (Kopp and Wachsmuth 2004; Leβmann *et al.* 2006) and *NUMACK* (Kopp *et al.* 2004).

Finally, we discuss which requirements a technical model must meet in order to be more compatible with GP theory. These include: (i) an analogical model of visuospatial as well as motor imagery; (ii) the representation of content in two distinct semiotic modes (that is, discrete categorial vs. analog continuous modes of meaning); (iii) a model of a dialectic for these modes; and (iv) the pervasive influence of discourse context on the form, timing, and content of speech-coexpressive gestures. We will outline how some of these requirements could be modeled computationally. In conclusion, we discuss what benefits can be expected for ECAs that conform to GP theory, in terms of "improved" communicative behavior, and we consider further implications of our results for modeling the comprehension of multimodal communicative behavior as well.

18.2 Growth point theory

McNeill (2005) presented a conception of language that acknowledges its dynamic dimension in an imagery–language dialectic, in which gestures provide the imagery. Gesture is an integral component of language, not merely an accompaniment or ornament. Such gestures are synchronous and coexpressive with speech, not redundant, and are not signs, salutes, or emblems. They are frequent—about 90% of spoken utterances in narrative discourse are accompanied by them (Nobe 2000).

18.2.1 Gesture and dialectic

The synchrony of speech forms and gestures creates the conditions for an imagery–language dialectic. A dialectic involves:

(a) conflict or opposition of some kind, and

(b) resolution of the conflict through further change or development.

The synchronous presence of unlike modes of cognition, imagery, and language, that are coexpressive of the same underlying thought unit, sets up an unstable confrontation of opposites. Even when the information content in speech and gesture is similar it is present in contrasting semiotic modes, and a dialectic occurs. This very instability fuels thinking-for-speaking (that is, thinking generated, as Slobin (1987) says, because of the requirements of a linguistic code) as it seeks resolution. Instability is an essential feature of the dialectic, and is a key to the dynamic dimension. The concept of an imagery–language dialectic extends a concept initiated (without reference to gesture) by Vygotsky, in the 1930s (cf. Vygotsky 1987):

> The relation of thought to word is not a thing but a process, a continual movement back and forth from thought to word and from word to thought. In that process, the relation of thought to word undergoes changes that themselves may be regarded as development in the functional sense. Thought is not merely expressed in words; it comes into existence through them.
>
> (Vygotsky 1987, p. 218)

This new conception also recaptures an insight lost for almost a century, that language requires two simultaneous modes of thought—what Saussure, in recently discovered notes composed around 1910, termed the "double essence" of language (although he expressed this without reference to gestures; cf. Harris 2002; Saussure 2002).

Gesture is naturally opposed to linguistic form. At the point where speech and gesture are synchronous they are coexpressive; they present the same underlying *idea unit* (an idea possessing possible internal complexity that functions as a single unit of meaning, attention, and memory) in two forms. The idea unit ties them together, and explains the synchrony. The opposition between them is semiotic, different ways of packaging information, and exists even when the referential content of speech and gesture is the same. In gesture, an idea unit is embodied *globally*, as a whole, instantaneously, and concentrates in one symbol what may be distributed across several surface elements of speech. Simultaneously, in speech, the same idea unit is represented *analytically*, *combinatorically,* and *linearly*. In this semiotic opposition the idea unit exists at the same moment in two semiotically opposite forms, a contrast that fuels thought and speech.

How is the form of a gesture determined by its meaning? Take the image of a wall: the wall has features, but they are not the origin of the image. The image is related to the context of speaking. If the locus of a wall is the significant point in the context, then perhaps no features of shape will be motivated at all, because a locus does not inevitably inhabit a shape; if the *field of oppositions* is such that verticality alone is the differentiating feature, this will be the image (and gesture); if horizontality is the differentiating feature, then that will be the image; and so forth. So features are a product of the differentiation of a global image, not the source, and are related to the context.

The instability or "tension" in a dialectic also requires a global perspective. The same idea unit is embodied in two opposed forms. This means that some way must exist to register the "sameness" of the idea, and also the opposition. The sameness is registered with respect to (a) differentiation from context and (b) reference; these dimensions are the same for both the linguistic and imagery components of the dialectic. The *instability* of the dialectic comes from the conflict of holding one idea in these two coexpressive modes. There is a third factor, a "force to resolve" the instability, without giving up either part. In nature (us), this is a specific case of homeostasis; for modeling it can be imagined as represented by an additional third force to resolve the opposition between the imagistic and linguistic modes. Such an approach may be useful in many ways, since it may open up experiments with the model in which the force setting is systematically varied, and the effects on resolving the dialectic observed.

18.2.2 Growth points

The smallest unit of the imagery–language dialectic is posited to be a "growth point" (GP), so named because it is, theoretically, the initial unit of thinking-for-speaking out of which a dynamic process of organization emerges. In the GP, interactions between spoken form and imagery occur continuously and in both directions, it is not that imagery is input to spoken form or spoken form to imagery; the effects are mutual.

A GP is an empirically recoverable idea unit (cf. McNeill 2005; appendix), inferred from speech–gesture synchrony and coexpressiveness. An example recorded in an experiment (offered in part because of its ordinariness) is a description by a speaker of a classic Tweety and Sylvester escapade, which went in part as follows: "and Tweety Bird runs and gets a bowling ba[ll and drops **it do**wn the drainpipe]".[1] Speech was accompanied by a gesture in which the two hands thrust downward at chest level, the palms curved and angled inward and downward, as if curved over the top of a large spherical object (Figure 18.1). At the left bracket, the hands started to move up from the speaker's lap to prepare for the downward thrust. Then her hands, at the very end of "drops," froze briefly in midair in the curved palm-down position (the first underlining). Next was the gesture stroke—the downward thrust itself—timed exactly with "it down" (boldface). Movement proper ceased in the middle of "down", the hands again holding in midair until the word was finished (the second underlining). Finally, the hands returned to rest (right bracket). The two pauses or holds and the continuing preparation phase itself reveal that the downward thrust was targeted precisely at the "it down" fragment: the downward thrust was withheld until the speech fragment could begin and was maintained, despite a lack of movement, until the fragment was completed. Significantly, even though the gesture depicted downward thrusting, the stroke bypassed the very verb that describes this motion, "drops," the preparation continuing right through it and even holding at the end.

The fragment, "it down", plus the image of a downward thrust, was the GP. It is impossible to fully understand the source of any GP without elaboration of its relationship to context. This relationship is mutually constitutive. A GP cannot exist without a context, because it is a point of differentiation within it; and the context is a representation created, in part, to make the differentiation possible. While context reflects the physical, social, and linguistic environment, it is also a mental phenomenon; the speaker *constructs* it in order to make the intended contrast, the GP, meaningful within it. Theoretically, a growth point is a psychological predicate in Vygotsky's (1987) sense (also Firbas 1971), a significant contrast within a context.

18.2.3 Gestural imagery

Even casual observation of gesticulating speakers reveals that gestures are often depictive of entities and events. Speakers are able with their gestures to *iconically* represent features of things that they have seen; for example, the flat sides of a box or the swift descent of a falling object. This makes it reasonable to suppose that gesture generation may be a straightforward process of transposing visuospatial imagery from a mental store to

[1] Notation: Square brackets [...] enclose the portion of speech that goes along with a *gesture phrase*, a sequence of movement phases containing exactly one *stroke* (Kendon 1980). The stroke is the meaning-bearing part of the gesture phrase, performed with effort, and the only movement phase that is obligatory. The opening bracket [marks the onset of the gesture phrase, when the hands start to move from rest or a previous gesture into position to perform the stroke;] is the end of the gesture phrase; boldface is the gesture stroke itself; underlining is a pre- or poststroke hold, a brief cessation of motion to ensure the synchrony of stroke and targeted speech.

Figure 18.1 Gesture stroke accompanying "it down" in the sentence "and drops it down the drainpipe". From McNeill (2005). Computer art in this figure by Fey Parrill.

speakers' hands and gesture space as they describe such objects and occurrences. Hadar and Butterworth (1997), proponents of such a view, state that gesture comes from visual imagery via a "direct route"; that it is "the motor manifestation of imagistic activation" (p.167). Certain comparisons among gestures produced in extended, narrative discourse contexts, however, reveal that there are factors in addition to mental imagery that motivate aspects of gesture form and execution. In the following discussion, we will focus on three such comparisons.

Figure 18.2 shows a sample of speakers who are individually telling, from memory, the "Tweety and Sylvester" cartoon story they have just seen. The three video stills on the right in Figure 18.2 are excerpted from descriptions of one cartoon event. This is an interval in the cartoon in which we see a cat climbing up a long drainpipe on the side of a building (as in the leftmost still in Figure 18.2). The cartoon cat's goal is to reach a bird

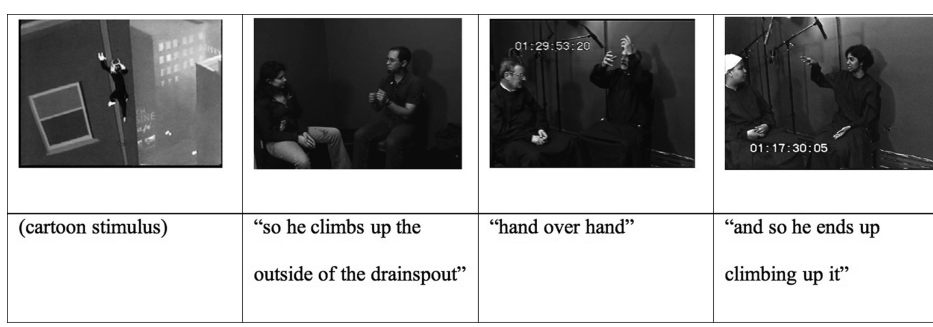


Figure 18.2 Three speakers' gestural depictions of a cat climbing up a drainpipe, as seen in the cartoon eliciting stimulus.

who is sitting in a window above. The interval of the cartoon is long enough for us to observe the cat's four legs moving alternately as he climbs up the pipe's length. Each of the three storytellers' descriptions of this event was contextualized in a sequence of recounted events comprising a 5- to 8-minute, continuous narration. Note that, despite having observed the same event in the cartoon, these speakers' gestures individually picked out somewhat different features of it for depiction, such that the gestures varied quite a lot in form and execution from one speaker to the next. In terms of the GP theory, such variations imply differences of thinking for speaking.

The leftmost speaker's closed fists move alternately upward a short distance, suggesting the cat's climbing manner of motion and his path upward. The second speaker's hands are open, appearing to grasp the virtual drainpipe. This speaker's gestured climbing motion extends a greater distance than the first speaker's, moving from abdomen-level to above his head. The third speaker represents the climbing manner of motion more abstractly, by simply wiggling the fingers of her right hand while moving it up to the level of her head. In other speakers' gestures, not shown here, climbing may not be represented at all, despite the way this feature of the cat's motion was made so noticeable in the cartoon. Some speakers, for instance, simply trace the cat's upward path of motion with an extended index finger.

The point we want to emphasize with this first comparison is that, despite having encoded the same visuospatial image from the cartoon eliciting stimulus, different speakers depict the features of that image quite variously. This amount of cross-speaker variability is a widely acknowledged characteristic of the unrehearsed, coverbal gesture that accompanies natural discourse. Given that all speakers can be assumed to have encoded the same image from the cartoon and all have the same "articulators" at their disposal (two hands, head and torso, gesture space), this variability suggests that visual imagery is not the sole determinant of gesture form and execution.

The video stills in Figure 18.3 demonstrate how an individual speaker's repeated references to a single witnessed event, across an extended interval of storytelling, are likely to be accompanied by gestural depictions of the event that differ in many features. In an elicitation similar to that represented in Figure 18.2, this speaker was telling the story of a short film about a man picking pears, some of whose pears are later stolen. The three

Figure 18.3 Three of one speaker's several gestures that accompany different mentions of picking pears, each differing substantially from the others in form and execution.

video stills on the right in Figure 18.3 are excerpted from the speaker's descriptions of one event in the film: the action of the man picking the pears. An interval in the early part of the film offers the viewer an extended close up of the man's hand grasping a pear and pulling it from the tree limb (the leftmost still). Across her extended narration, the storyteller refers to this action of pear-picking five different times. At each mention, the co-occurring gesture is different in form and execution from all other mentions.

Three of these gestures accompanying mentions of pear-picking are shown in Figure 18.3. At the speaker's initial mention of the activity—"he's picking the pears"—her left arm is raised and she makes a grasping motion with that hand. This is an iconic representation of picking a pear from a tree limb, high up. With her left hand she "pantomimes" plucking a pear off the limb,[2] a gestural image with an easy-to-perceive resemblance to the eliciting video image. At the next mention, 10 to 15 seconds later in her narration (not shown), she performs another left-handed gesture, similar in form and execution, but larger and repeating. By the time she mentions pear picking again, about 45 seconds later, the speaker has described other events in the story line that occur while the man is occupied up in the tree. At this third mention, she performs a metaphorically representational gesture that does not depict pear picking in any way. As she says, "he's picking his pears," both hands are held out as if presenting something. This is a reintroduction of given information, abstractly, as a discourse entity: the pear picking that continues in the tree above the narrated activities that are occurring meanwhile on the ground. This final mention of picking pears occurs near the end, at a point in the story line where the man in the tree again becomes an object of focus. Just prior to saying, "picking the pears," she performs a gesture with her right hand that represents the path of motion; specifically, the trajectory of some other characters who walk past the man's pear tree. Superimposed on this path gesture, the right hand briefly makes a flapping movement that is only minimally suggestive of removing pears from the tree.

This second comparison underscores the point of the earlier comparison of different speakers' gestures; that is, that recalled visual images are not the only factor motivating features of gesture form and execution. This second comparison shows this to be true within speaker and within a single coherent story. Each mention of pear picking is with a different discourse purpose. These different purposes, together with details such as which hand was currently engaged in gesturing and particulars of the immediate story context, exerted a shaping pressure on gesture form and execution. The impact of such factors extended to choice of hand, how large and feature-rich or pantomimic the gesture would be, and where in gesture space the gesture would occur. In parallel with our preceding interpretation of cross-speaker variations as indicative of slight differences

[2] Pantomime is gesture without speech often in sequences and usually comprised of simulated actions. With gesticulation the individual speaker constructs a combination of speech and gesture, combined at the point of maximal coexpressiveness. In pantomime, none of this occurs. There is no coconstruction with speech, no coexpressiveness, and timing is different, if there is speech at all. The very same movement—that in Figure 18.1, for example—may occur as a pantomime or as a gesticulation. Whether the speaker combines such movement with speech is the key discriminating factor.

of thinking for speaking, here we see the development of *intra*-speaker variations—the gestural image of pear picking shaped within constantly shifting contexts as the discourse is built up.

18.2.4 Gesture and speech synchrony

Our third comparison of gestures starts by expanding on the theme of how discourse processes exert a shaping pressure on gestures. In a recent study, Duncan and Loehr (in preparation) explored the impact of the changing "contrastive discourse focus", a narrative cohesion phenomenon, on how visual imagery manifests in gesture. Speakers' renditions of two events that occur about one minute apart from each other in the cartoon story referenced in Figure 18.1 were compared. The events are similar in that each involves the cat climbing up the drainpipe on the side of the building in order to reach the bird above. In the first target event, the cat's initial act of climbing, he climbs up on the outside of the drainpipe. The second target event, his second act of climbing, is via the inside of the same drainpipe.

From observations of many full-length (5- to 8-minute) cartoon narrations, we find that speakers typically make the feature of "inside" versus "outside" a point of contrastive discourse focus in their descriptions of these two target events. When describing the cat's first ascent, they will often just say, "he climbed the drainpipe". Or they may say, "he climbed up outside the drainpipe". In these utterances, typically, speech prosodic emphasis is given to the verb "climb". Subsequently, with great regularity, when describing the cat's second ascent, speakers choose words and a prosodic intonational contour for their utterances that together emphasize the "inside" aspect of the ascent; for example, "this time he climbed up INSIDE the drainpipe". In other words, the content that is contrastively discourse focal, the new information that differentiates one utterance from the last, is given prominence through word choice and speech prosodic emphasis.

With respect to the gestures that accompany such utterances, we find that whatever event feature is contrastively focal in the discourse at the moment; this feature is also typically the substance of coverbal gesturing. Figure 18.4 shows one speaker's descriptions of these two target cartoon events. The top row shows video stills from the two target cartoon events; in the bottom row, video stills from a speaker's narration of the respective event along with transcripts of the speech are shown (capital letters identify the interval of speech that is given prosodic emphasis, that is heightened pitch and increased loudness and syllable length; see Loehr and Duncan, in preparation). As the speaker describes the cat's first ascent, his closed fists move alternately upward, suggesting the cat's climbing manner of motion and his path upward. This manner-expressive gesture is performed twice, first in synchrony with the manner-expressive phrase "climb up the", and then in synchrony with the also ground-expressive (the pipe) phrase "climbs up the outside". So, we see synchronized coexpression of semantic content in the two modalities.

Similarly, about a minute later, when describing the cat's second ascent of the drainpipe, gesture and speech are coexpressive. This time, however, the act of climbing is not the focus of the speaker's discourse. Even though the speaker still conceives of the cat as climbing (he repeats the verb "climb") his gesture does not show climbing manner,

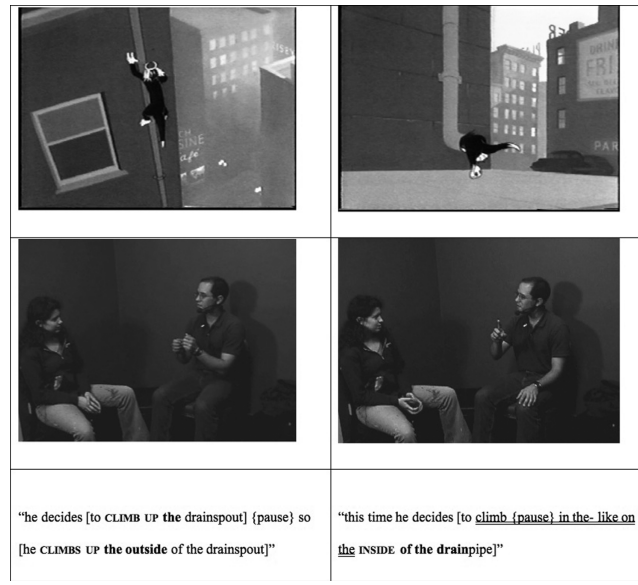


Figure 18.4 One speaker's descriptions of the cartoon cat's two attempts to ascend the drainpipe, first on the outside, then on the inside of the pipe.

nor does it synchronize with that verb in the utterance. The stroke phase of this gesture is the speaker's extended index finger pointing and moving along a path away from his body and then upward. This stroke phase skips the verb "climb" to synchronize instead with "inside", the figure-ground relational term that captures the contrastively focal element at this moment of the discourse.

Significant for our understanding of the relationship between visuospatial imagery and gesture form and execution is the fact that, prior to performing this stroke phase there is a prestroke hold phase beginning with the verb "climb" and extending across several words. This means that the speaker had the makings of his gestural image ready to produce by the time he finished uttering the words, "he decides to". However, as we regularly see in this discourse context across many cartoon narrations, the speaker held this gesture, waiting until the element of speech that would coexpress the contrastively discourse focal feature of "inside-ness" arrived in the sequential speech stream.

Gestures can densely encode many features of the entities and events that people image in their minds as they speak. That just as many gestures are quite reduced, quite selective in the features of visuospatial imagery they express, reveals the working of constraints that, according to GP theory, are part of the language production process itself. Our comparisons give clues to the nature of at least some of the constraints during a storytelling-type discourse. Speech and gesture coexpress in a very tight synchrony the contrastively discourse focal elements of information. The temporal synchrony is ensured by active pre- and poststroke hold phases. The coexpressivity results from the fact that, as the outside–inside comparison reveals, both modalities coordinate to highlight those features of complex events that are most focal in the speaker's thinking at

the moment of producing a discourse-embedded utterance. The focal center of each such utterance is the element of new information that contrasts with background elements built up in the preceding discourse. Joint highlighting of this element by the two modalities, serves (we assume) to focus both speaker and recipient attention on the information that propels the discourse forward toward the narrative goal the speaker has in mind at the moment. This overview shows that discourse context is a factor determining gesture form and execution in relation to co-occurring speech. Gestures, rather than coming from visual imagery via "direct route" (Hadar and Butterworth 1997), are revealed by the representative examples discussed above to be discourse-embedded, linguistic-conceptual representations (McNeill and Duncan 2000; see also, Duncan 2002) whose form is dependent upon the speaker's discourse focus of the moment. This fact has clear implications for efforts at modeling speaking-associated gesture in an ECA. Before entering this discussion, we review the assumptions and models adopted in current ECAs.

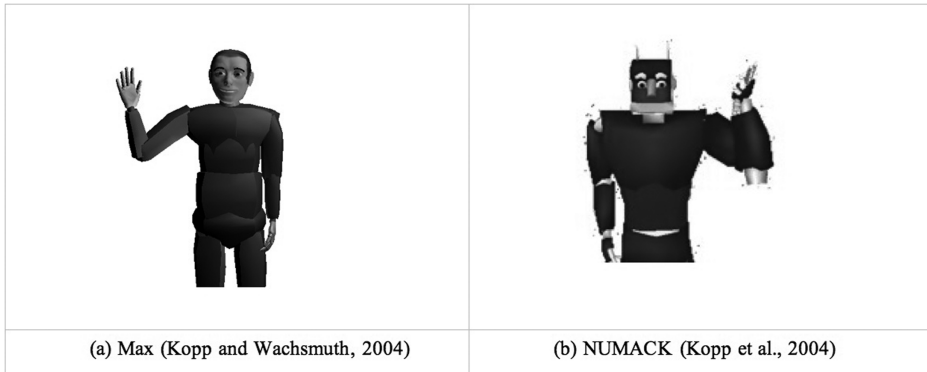
18.3 Gesture and speech in embodied conversational agents

ECAs are computer-simulated characters that possess much of the same overt interaction abilities as humans demonstrate in face-to-face conversations. This involves the production of utterances that are composed of simultaneous and synchronized verbal and nonverbal behaviors. So far, one main challenge in building ECAs has been to automatize the generation of natural-looking multimodal output, without entirely relying on static, predefined, and thus limited, repositories of canned behaviors. A generation model that comes anywhere close to the generative power of humans' speech and gesture performances requires a time-critical production process with high flexibility.

In technical approaches, and contrary to GP theory, this process has been conceived of in terms of modular stages that more or less directly correspond to the stages assumed for Natural Language Generation (e.g. Reiter and Dale 2000). These architectures are construed as modular, pipeline models broken down into three subtasks—content planning (also known as text or document planning), behavior planning (microplanning), and behavior realization. Starting from a goal the speaker wants to achieve, in ordinary language, the work done by these three subsystems may be summarized as: figuring out what to say, figuring out how to say it, and, finally, saying it. These stages are crucially linked to each other and must operate not only on speech but include other modalities like gesture as well. In this section we review two state-of-the-art ECAs, Max and NUMACK (Figure 18.5), which focus on the latter two stages, behavior planning and behavior realization, which coarsely correspond to the cognitive processes that GP theory aims to explain. In contrast to other ECA implementations, the approach used in Max and NUMACK is an attempt to generating coordinated gesture and speech on-line.

18.3.1 Behavior realization in Max

Behavior realization concerns the ability to generate various verbal and gestural behaviors in real-time, from some sort of representation that specifies the decisive features of



(a) Max (Kopp and Wachsmuth, 2004) (b) NUMACK (Kopp et al., 2004)

Figure 18.5 (A) Max (Kopp and Wachsmuth 2004) and (B) NUMACK (Kopp *et al*. 2004), two ECAs that embody models of speech and gesture production.

these behaviors and the temporal relations between them. In the virtual human Max, the "Articulated Communicator Engine" (ACE, for short) is employed for this task. ACE is a software platform that allows one to create and visualize animated agents, and to synthesize for them multimodal utterances including speech, gesture, or facial expression. Input descriptions are formulated in MURML, an XML language for succinctly defining multimodal behavior (Kopp and Wachsmuth 2004).

The ACE production model aims at creating lifelike, synchronized verbal and non-verbal behaviors in a human-like flow of multimodal behavior. To this end, it tries to simulate the main mutual adaptations that appear to take place between speech and gesture, when humans try to achieve synchrony between the coexpressive elements in both modalities. One hallmark of the ACE approach is an incremental process model that allows for handling cross-modal interactions at different levels of an utterance, corresponding to decisive points in multimodal behavior generation. In accordance with GP theory, the ACE production model is based upon an empirically suggested segmentation hypothesis (McNeill 1992), that continuous speech and gesture are coproduced in successive multimodal *chunks* each expressing a single idea unit. The incrementality of speech–gesture production is reflected in the hierarchical structures of overt gesture and speech and their cross-modal correspondences. Kendon (1980) defined units of gestural movement to consist of gesture phrases (cf. Footnote 1 for further explanations). Similarly, the phonological structure of connected speech in intonation languages such as English and German is organized over intonation phrases (e.g. cf. Levelt 1989). Such phrases are separated by significant pauses, they follow more the semantical (deep clause) structure than the syntactical phrase structure, and they have a meaningful pitch contour with exactly one primary pitch accent (the nucleus).

ACE takes chunks of speech–gesture realization, as produced in trouble-free utterance, to be pairs of an intonation phrase and a coexpressive gesture phrase (see Bergmann and Kopp 2005 for empirical evidence for this). That is, complex utterances with different

gestures are considered to consist of several chunks, with the aforementioned synchrony holding within each of them. While GP theory assumes that temporal synchrony between coexpressive speech and gesture is inherent to the dialectic in which they come to exist, ACE tries to produce these elements in synchrony by utilizing adaptations between speech and gesture. Based on the segmentation hypothesis, cross-modal adaptations take effect either within a chunk or between two successive chunks.

Within a chunk, temporal synchrony between certain words and the stroke is mainly accomplished by the gesture's adapting to the timing of speech, while speech runs mostly unaffected by gesture ("ballistically"). In producing a single chunk, the intonation phrase is therefore synthesized in advance, possibly augmented with a strong pitch accent for narrowed focus. As in related systems, ACE exploits information about absolute phoneme timings retrieved from a text-to-speech system (TTS) to set up timing constraints for coverbal gestural or facial behaviors. The gesture stroke is thereby set either to precede the coexpressive speech's onset by a given offset or to start exactly at the nucleus (the most prominent pitch accent) if a narrow focus has been applied. Further, the stroke is set to span the whole portion of speech that is associated with the gesture (its *lexical affiliate*) before retraction starts. This is achieved either by inserting a post-stroke hold after a normally executed stroke phase, or by performing additional repetitions of the stroke.

Humans often anticipate the synchrony between speech and gesture *before* the next chunk starts and adapt their speech and movements accordingly. ACE reproduces main preparatory effects in both speech and gesture, taking place at the boundary between two successive chunks. First, the onset of the gesture phrase covaries with the position of the nucleus and, secondly, the onset of the intonation phrase covaries with the stroke onset (de Ruiter 2000; Nobe 2000; McNeill 1992). In consequence, movement between two strokes depends on the timing of the successive strokes and may range from the adoption of intermediate rest positions to direct transitional movements (so-called "coarticulation effects"). Likewise, the duration of the silent pause between two intonation phrases may vary according to the required duration of the preparation for the next gesture. ACE simulates these highly context-dependent adaptation effects during the phase when the next chunk is ready to be uttered ("lurking") and the preceding chunk is "subsiding," that is done with executing its meaning-bearing parts (intonation phrase and gesture stroke). It is at this time when intrachunk synchrony is defined and reconciled with the onsets of the phonation and the preparation, and that all gesture animations are created such that they satisfy the movement and timing constraints now determined.

For example, suppose that Max has just completely uttered the intonation phrase of a chunk, has performed the corresponding gesture stroke, and is now moving his hands back to a rest position. In the next chunk, which belongs to the same utterance and thus is to be seamlessly connected, the linguistic elements that are coexpressive with a gesture are located relatively early in the intonation phrase, and the gesture requires—under current movement conditions—an extensive preparation. Thus, movement needs to start early in order to meet within realistic speed constraints the mandatory timing of

stroke onset. ACE will create a fluent gesture transition after an only partial retraction, according to the position of the coexpressive speech within the next verbal phrase. In turn, the vocal pause between the intonation phrases is stretched as needed for the speech-preceding preparatory movement.

The ACE process model exceeds other systems, for example the BEAT (Behavior Expression Animation Toolkit) system (Cassell *et al.* 2001), in that it enables speech and gesture to interact and to coordinate with each other *during* the uttering of a multimodal chunk. Still, speech and the single phases of a gesture are executed as preplanned in a feed-forward manner. While this allows for an exactly timed gesture stroke, possibly extended with a poststroke hold, this level of interactivity is still insufficient for simulating prestroke holds as described in Section 18.0. It is conceivable that the motor control layer of ACE can be utilized to enable prestroke holds, notably, by constructing two distinct sets of local motor programs (LMPs) for the preparation and the stroke up front. Planning these LMPs sets to blend smoothly per default ensures a continuous entry into the gesture stroke; predicating the initiation of all stroke LMPs upon the arrival at the coexpressive verbal elements results in the emergence of finely adapted prestroke holds. The main problem, then, is to stream synthesized speech in a way that allows monitoring of the appearance of distinct points.

18.3.2 Behavior planning in NUMACK

The conception and real-time capable implementation of models for behavior planning, that is the problem of determining coordinated language and gesture forms, is one of the hardest challenges ECA research is facing. Previous systems, in particular REA (Real Estate Agent) (Cassell *et al.* 2000b), extended a natural language grammar formalism to handle constituents to be uttered in different modalities. REA was able to generate gestures, and to coordinate them with the meaning of the linguistic expression they accompany and the discourse context within which they occur. However, whole gestures were lexicalized like words, selected using a lexical choice algorithm and incorporated directly into sentence planning. While this approach allows for context-dependent coordination with speech, it does not allow for the natural generative power of gestures that form to express new content.

The NUMACK system (Kopp *et al.* 2004) has tackled the *formation* of iconic gestures based on systematic meaning-form mappings. This approach is based on the assumption that iconic gestures communicate mainly in virtue of their resemblance to visuospatial properties of the entity they depict. Even if an iconic gesture may by itself not uniquely identify an entity or action in the world, it always depicts (or specifies) features of an image through some visual or spatial resemblance. To account for how iconic gestures are able to express meaning, this work provided a way to link gestures to their referents by assuming an intermediate level of abstraction and representation that accounts for a context-independent level of visual-spatial meaning.

Generation of iconic gestures in NUMACK was based on the view that, if iconic gestures are communicative of such imagistic information, and if people are able to recover and interpret this meaning, there must be a reliable system of ways of depicting

imagistic content. The hypothesis, thus, was that there are prevalent patterns in the ways the hands and arms are used to create iconic gesture images of the salient, visual aspects of objects or events, and that such patterns may account for the ways human speakers derive novel gestures for objects they are describing for the first time. Furthermore, the generativity that human gesture displays was taken to suggest that such patterning or commonality pertains not to the level of gestures as a whole, but to subparts—features of shape, spatial properties, or spatial relationships that are associated with more primitive form features of gesture morphology, like hand shapes, orientations, locations, movements in space, or combinations thereof.

Based on these assumptions, a feature-based approach was adopted in the NUMACK system to model the intermediate level of meaning that links gesture to the imagistic content it depicts. Separable, qualitative *image description features* (henceforth, IDFs) were used to describe the meaningful geometric and spatial features of both a gesture's morphology and the entities to which a gesture can refer. It was further assumed that iconic gestures are composed of sets of one or more morphological features that convey sets of one or more image description features, and that each of these mappings from IDFs onto form features can be found in different gestures depicting different, but visually similar, things. This level of granularity allowed for explaining and modeling how gestures can communicate, without having standards of form or consistent form–meaning pairings.

Consequently, behavior planning in the NUMACK system comprised a gesture planner that is responsible for planning a gesture morphology appropriate to encode a set of one or more input IDFs. Similar to a sentence planner for language, the gesture planner drew upon an input specification of domain knowledge, plus a set of entries to encode the connection between semantic content and form. Form–meaning coupling were formalized in a set of "form feature entries", data structures that connect (conjunctions of) IDFs to (combinations of) morphological features. When receiving a set of IDFs as input, the gesture planner searches for all combinations of form feature entries that can realize them, and combines them by iteratively filling a morphology feature structure for a gesture. That way, the gesture planner builds up gestures until as much as possible of the desired communicative effects are encoded. Assume, for instance, that an input IDF representing the "verticality" of an object (e.g. a tall landmark in a route description) is to be expressed by a gesture. The system will retrieve the form feature entries connected to the IDF "verticality" and choose one of several morphological features able to express this piece of content—for instance a flat hand pointing upwards. Additional IDFs may be encoded iteratively taking the constraints and limited degrees of freedom imposed by the choice of the "flat hand upwards" feature into account. Note that the system may output an underspecified gesture if a morphological form feature does not meaningfully correspond to any of the disposed IDFs, that is it remains undefined by the selected patterns. As a result, the gesture planner will provide a set of gestures, each of which is annotated with the IDFs it encodes. Based on this information, the sentence planner combines them with words in predefined structural ways, in order to derive full multimodal utterances. Resulting utterances are then passed on to ACE for on-demand realization.

18.4 Modeling challenges and possible solutions

GP theory and its empirical foundation as sketched in Section 18.0 imply certain properties for a computational model that we will discuss in the following. We will contrast these properties with the gesture production approaches followed with the Max/NUMACK systems as described in Section 18.0. Then, we discuss possible solutions to be found in a computational model that is more compatible with, and capitalizes on, GP theory and its empirical underpinnings.

18.4.1 Problems with features

Current ECAs like Max or NUMACK work from stored feature decompositions of objects to compound gesture forms. This means, features must be present *a priori* for the mechanism to work. Such an approach touches upon the question of whether one can assume features or combinations of features to have morpheme-like properties, that is, that they are meaningful pieces that cannot be divided into smaller meaningful parts but that can be combined according to certain rules to compose larger units (gestures). According to GP theory, the features of human gesture are dependent on meaning, arise out of global imagery, and do not exist *a priori*. Indeed, spontaneous gestures do not have standards of form, but under conditions that we are just beginning to study they can develop a degree of *form stability* via features and correlated meanings. Nonetheless, even after stabilization there are inconsistencies incompatible with standards of well-formedness. More importantly, the sorts of gestures that are the focus of our modeling efforts here lack essential morphemic characteristics.

In spontaneous gestures produced for an audience in a context where speech is not allowed we can find examples, when the same gesture is repeated, where the form stabilizes, maintains distinctiveness *vis-à-vis* other forms, and undergoes morphological simplifications that appear to maintain distinctiveness with increases in fluency. In the "Snow White" corpus (see McNeill 1992, pp. 65–72), for instance, a subject is retelling this fairytale exclusively with gestures, no speech allowed. Two gestures (King, Queen) contrasted immediately and showed substantial changes as they were used. These changes increased fluency, but the Queen–King contrast remains stable. The "listener" adopts this gesture system and even a conversation using only the newly established morphemes can be observed. Thus, ritualization is apparent in these gestures. Looking at the Canary Row (Tweety and Sylvester) corpus, however, we see factors that promote stabilization of gesture forms. But these resulting gestures lack any kind of simplifications or distinctiveness *vis-à-vis* other gestures, arguing against a morphological structure. Hence, we cannot assume that the mere presence of an audience is sufficient for a stabilization that may cause true morphemes to develop.

The importance of this question is clear in Max. A feature vocabulary can underlie both his production and perceptual processes when the domain of discourse is restricted, for example to the shapes of virtual objects. Then the features of these objects are known in advance and can be listed with correlated possible meanings—a kind of morphology. But in the process of GP formation features are emergent in most cases.

18.4.2 An analogical model of visuospatial and motor imagery

From the point of view of GP theory, gesture is considered as embodied visuospatial and motor imagery. A direct consequence of this assumption for a computational approach is that gesture should arise and develop from the activation of imagery, which is not produced from an internal imagistic representation by some sort of symbolic transduction process (cf. Barsalou 1999). This *embodied imagery* hypothesis of GP theory blurs the distinction between content representation and processing or action execution, a concept that is virtually foundational for computational modeling except in connectionist approaches, and that is also the foundation for the gesture generation pipeline in ECAs (cf. Section 18.0). The "motor units" responsible for limb movement would no longer be just executing modules but, at the same time, are also representational units for imagery. This approach contrasts with the "pipeline" production in ECAs that distinguishes content planning, behavior planning, and behavior realization. This three-stage, modular approach detaches content (imagery) from gesture (motor).

Three "types" of imagery can be observed in gestures, all of which need to find their equivalent in a model. In narrative discourse one can usually find two viewpoints from which gestures are produced (McNeill 2005). In character-viewpoint (C-VPT) gestures the hands or the body of the speaker represent corresponding body parts of a character in a narration, while in observer-viewpoint (O-VPT) gestures the hands represent entities in the narration. Thus, C-VPT gestures in narrations embody motor imagery, bodily action of another character mapped onto one's own body, and O-VPT gestures embody visual or spatial imagery. Though there is no clear border between visual and spatial imagery, the former term emphasizes the imagination of visual appearance that may give rise to gesture (e.g. outlining an object in two dimensions), while the latter emphasizes aspects of spatial configuration and layout, not necessarily experienced visually. Spatial imagery can be frequently observed in gestures for route descriptions, for places, or for complex objects. A characteristic property of gestures in these domains seems to be spatial cohesion, that is the creation of a complex image spanning multiple, successive gestures (Emmorey *et al.* 2000; Enfield 2004).

What does "imagery" look like in computational models? A prominent modeling approach for spatial imagery is to use two-dimensional, matrix-like structures that represent an analogical spatial layout for relational information, for instance in verbal expression. Glasgow (1993) describes an implementation using symbolic arrays in which neighboring cells analogically represent neighboring areas of (two dimensional) space, such that relative spatial and topological relations are implicitly represented. The cells are occupied by symbols that represent entities. The spatial representation of a proposition such as "*the spoon is to the left of the knife*" would be an array in which a cell occupied with the symbol *spoon* is left of a cell containing the symbol *knife.* Though this type of symbolic array is, according to Glasgow, no more expressive than a propositional, logic-based representation, the symbol array was shown to be much more computationally efficient with respect to typical spatial inferences. Besides a spatial representation, Glasgow's model also incorporates a visual component, implemented with three-dimension occupancy arrays that approximate object shape. Kosslyn (1980, 1987) suggests two-dimension matrix

structures for the representation of an object's visual appearance. In his model, such structures exist for long-term storage as well as for working memory. Short-term "surface images" are manipulated in a visual buffer consisting of retinotopically arranged points or "pixels". The same matrix structure is used for long-term storage of visual information, called *literal encodings*. These long-term representations are hierarchically structured such that a coarse skeletal encoding defines shape in a first approximation. Additional encodings for local regions or parts may elaborate this description. Global and local encodings are connected via spatial relations modeling the spatial layout.

A model of computational imagery in all three spatial dimensions is suggested by Croft and Thagard (2002). It is based on a scene graph, a representational structure used in computer graphics. A scene graph represents an object or multiobject scene as a tree structure with geometrical primitives—usually represented as triangle or polygonal surfaces—at the leaf nodes and geometrical transformations (among other properties) at the intermediate nodes. The transformations determine the spatial relations between the primitives that compose the object or scene. Scene graphs thus combine the visual and spatial components of imagery. Sowa and Wachsmuth (Sowa and Wachsmuth 2005; Sowa 2006) describe a model of visuospatial imagery, called *Imagistic Description Tree* (henceforth, IDT), developed to capture the imagistic content of shape-related gestures in a gesture interpretation system. Though structurally similar to a scene graph approach with a hierarchy of geometrical transformations, the terminal nodes in an IDT do not represent geometries, but coarse, qualitative specifications of shape in terms of an object's spatial extent—not unlike the IDFs used for gesture production (cf. Section 18.0)—and the qualitative course of its boundary.

In the light of the discussion about imagery (cf. Section 18.0) in which we pointed out the diversity of communicative aspects embodied in gestures, models of visual imagery based on two-dimensional "pixel" images or three-dimensions occupancy grids appear inappropriate for gesture production. These models too narrowly focus on visual appearance while they lack a potential for abstraction. It is, for instance, barely imaginable how the property of "being inside" can be represented with a pixel image that only captures this single aspect while omitting everything else. The semantic features of the IDT model and the IDFs that allow representing imagery in an abstract way are a step into this direction, yet they are confined to special semantic domains. A much greater variety of semantic primitives would have to be implemented in a computational model. As for the representation of spatial imagery it is at least evident that we need some model for spatial configurations, because successive gestures often use space in a cohesive way. Yet, it is not clear whether a qualitative approach (e.g. symbolic arrays) suffices, or a quantitative representation (e.g. scene graphs) is preferable, until the nature of spatial cohesion in gestures is examined in some more detail.

One, biologically inspired, way to implement a direct link between visuospatial imagery and motor processes is via association and spreading activation—a standard technique in associative networks. Let us assume a single (but structured) network of interconnected units. Each unit possesses a state of activation and influences (i.e. stimulates or inhibits) other units. The overall state of activation of the whole network "represents" imagery.

Some of the units are responsible for action execution, thus they can be considered the system's motor imagery. The activation of visuospatial imagery directly causes activation of motor imagery. Hence, both are integral parts of one single system, without any symbolic translations in between. Further, such an associative network, in which activation flows in any direction, merges the *representation* and the *processing* of imagery in a single system. A full-blown, network-based model of imagery is yet to be technically tackled, as is the then subsequent challenge of, for instance, connecting it to representation of linguistic meaning in a way that reconstructs a field of semiotic oppositions.

18.4.3 The global-synthetic property

Above, we pointed out the necessity for a GP-compliant computational model to reflect the global-synthetic property of coverbal gestures. This requirement conflicts with the bottom-up, feature-based approach currently used in computational architectures for the production of coverbal gestures by ECAs. Instead of constructing a gesture in a component-wise manner based on features with decontextualized meanings, in a GP-style solution a gesture would embody meaning as a holistic unit. Thus, the gesture as a whole is primary, while form features and corresponding meanings are secondary attributions by an observer.

The "global" requirement implies a computational model that produces gestures by means of *specialization* of a rather unspecific, schematic movement to a motor action that embodies all significant aspects (deviations from immediate context) at the same time. In order to implement a specialization approach in a computational system, we suggest organizing these unspecific movements in templates or "coordinative structures" (Kelso *et al.* 1983). Coordinative structures, a term coined in biology, functionally bundle different kinds of complex actions and explain how the motor system jointly coordinates several degrees of freedom in complex movements involving multiply redundant muscles and joints. They are functional groupings of different muscles working together to achieve a behaviorally relevant movement goal and controlled by fewer, abstract parameters.

In Saltzman and Kelso (1987), the movement goal was conceived of as an attractor of a set of differential equations with different types of underlying dynamics for different kinds of movement. Non-repetitive reaching movements, for instance, could be modeled by a dynamic mass-spring system with a point-attractor topology. The attractor in this case represents the final destination of the end-effector (the hand) in a reaching movement. The differential equations "pull" the end-effector to its destination regardless of the initial state or any perturbations during movement execution. For the modeling of repetitive or oscillatory movements, such as a circular movement of the hand, a periodic attractor topology was suggested. This basic approach could be extended to more complex movements. In the case of grasping, for instance, the attractor would be the affordances of a real object.

If we adopt the idea of coordinative structures for gestures, the movement goal, possibly modeled by the attractor, would be the significance. During gesture production a coordinative structure is shaped by significance, and thereby acquires meaning. The coordinative

structures zero in on these attractors; the properties of the attractor bring out features of gesture form in the coordinate structures interactively: so features are outcomes, not initial conditions, with significances that derive from the action as a whole, and this is the global property. There is no lexicon of feature–meaning pairs, but the features arise during the action itself. Once a gesture has been created it is usually true that we can identify features of form that carry meanings, but these would be the outcomes of the gesture, not the source. Each coordinative structure is an "action primitive", but without having significances by itself.

What kinds of coordinative structures, movement, or action primitives can we conceive of for gesture production? Some researchers consider gesture to be implicit action, derived from everyday practical experience. Müller (1998) distinguished "drawing", "outlining", "sculpting", and "grasping" as the basic action patterns expressed in depictive or iconic gestures. In a similar vein, Streeck (unpublished) suggests seven gesture practices: setting of gesture spaces (as a kind of model building), shaping (working on an imaginary substance), motion depiction (both real and fictive), haptic depiction (as handling objects), remote indexing, and mimesis (imitating bodily action up to re-enacting experienced scenes). Each of these general strategies for action-related gestural movement could be reflected in a computational model by a flexible action primitive or coordinative structure.

In order to use action primitives or templates for gesture generation in ECAs, two main problems have to be solved: (1) how do the primitives come about and (2) how do "meaningless" primitives connect to significance such that a concrete instance of the template embodies the targeted meaning?

If gestures effectively derive from practical action, building up a library of gesture templates should be a process following and depending on the ontogenetic acquisition of motor behavior. An appropriate computational model of that process in an ECA would thus require the agent to learn how to act in the world, and how to perform goal-directed actions on objects. However, since a virtual agent cannot have practical experience in the real world, a repertoire of behaviorally relevant actions has to be acquired by other means. One possibility could be imitation learning which has been successfully applied in robotics (cf. Billard 2002 for an overview). Kopp and Graeser (2006) suggested imitation learning for the acquisition of gestural motor behavior in ECAs (see also Kopp *et al.*, this volume). Their approach is based on motor command graphs that incorporate the agent's repertoire of motor commands given a context, and the position of the agent's body parts in space. Using the learning system, a virtual agent is able to immediately imitate known motor sequences, and to extend his motor repertoire if observed behavior (e.g. by another virtual human) does not match any known movement sequence. Such motor control graphs, acquired by learning via observation, could play the role of gesture templates if they are sufficiently abstract to represent a class of gestures derived from practical action (e.g. grasping).

In order to connect templates and significance, both have to be parameterized such that the free parameters of a gesture template can approximate the parameters of significance ("meaning shapes the utterance"). A general schema for stylized grasping,

for instance, will be used to create an iconic gesture accompanying an utterance like "he's picking the pears". If the idea of "pear picking" is the new contribution to the discourse context, it is likely that the verbal utterance is accompanied by a gesture. In that case, some parameters of the significance ("pear picking") influence parameters of the gesture template for stylized grasping resulting in the depiction of a grip appropriate for the size or shape of a pear. Associative learning by demonstration could be employed to associate the two sets of parameters. Given a meaning, for instance the shape of an object in a suitable parametrized representation, a human demonstrator could perform an appropriate gesture which is recorded by the system. The system will then associate the movement parameters of the gesture template with the parameters of the semantic representation. Using this learning paradigm, a static one-to-one mapping of meaning features on form features can be avoided.

Figure 18.6 illustrates a hypothetical example of "global" gesture learning. Here the significance (right side) is the cylindrical shape of an object and this meaning is expressed with a "three-dimentional-sculpting" gesture (left side). We assume that a parameterized schema for two-handed gestural sculpting, for instance as a coordinative structure or a motor control graph, exists or was acquired via imitation learning. In order to apply associative learning, the gesture schema has to have the ability to produce an action sequence, and to recognize the sequences it may produce. In the example, the gesture schema has two numerical parameters, movement length and hand distance. We further assume that a suitable parameterized representation of the cylinder's shape exists. Here we assume the extents of the major and minor perceptual axes to be numerical parameters (cf. Sowa and Wachsmuth 2005; Sowa 2006). Both schemas adjust their parameters appropriately upon perception. When gesture and object are presented at the same time during a training phase, associative links between their parameters are built or amplified if they contain similar values, and diminish if dissimilar. For the sculpting gesture in the example a large numerical movement length most probably corresponds to a large extent of the major object axis such that the association between these two

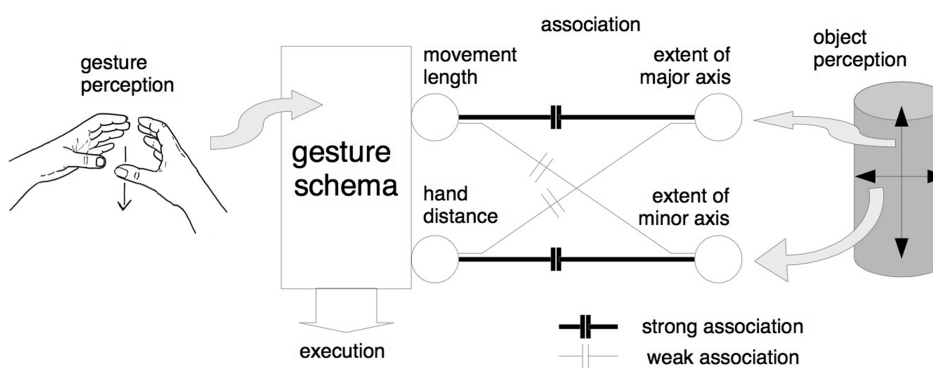


Figure 18.6 Learning gesture schemas: strong covariation of gesture form and object variables results in strong associative links between the variables.

parameters is amplified. Similarly, hand distance and the extent of the minor axis are likely to correspond and build strong associative links.

After learning, gestures are produced by activating "significances" (see Figure 18.7). Activations of the meaning parameters will spread to the parameters of one or more gesture schemas and eventually lead to the execution of a gesture. In contrast to the "constructive" approach associative learning does not rely on a fixed 1:1 mapping between individual form and meaning features and thus comes closer to a "global" gesture generation approach. It allows a specialization of a generic, non-meaningful gestural action represented in a schema via properties of the significance that "shape" the action.

18.4.4 Representation of content in two distinct semiotic modes and the dialectic

While gesture schemas in the form of coordinative structures together with a training procedure may partially model the *global* property of gestures, they do not model the GP itself: the differentiation of psychological predicates, growth, inseparability from context, copresence of imagery and linguistic categorization, the coexpressiveness of imagery and language, internal tension and motivation, or change/unpacking. In short, the *essential duality* of language of which the GP is a minimal unit, seems at present impossible to model by a computational system.

"Growth" in the sense of GP theory is a loose concept that defies definition. It includes the idea that new structure emerges out of old with a connection in between, but how to model this in a dialectic-type process is unclear. Dynamically speaking, the GP attracts effort, and this is realized as prosodic peaking on a linguistic segment and in the gesture (prosody being affiliated with gesture); it also becomes the focus of consciousness, of instantaneous being, and the "L-center" (the locus in speech of focal awareness, akin to the concept of a P-center, the point of focus in perception). The unpacking by a grammatical construction, on the other hand, is penumbral and supportive. Theoretically,

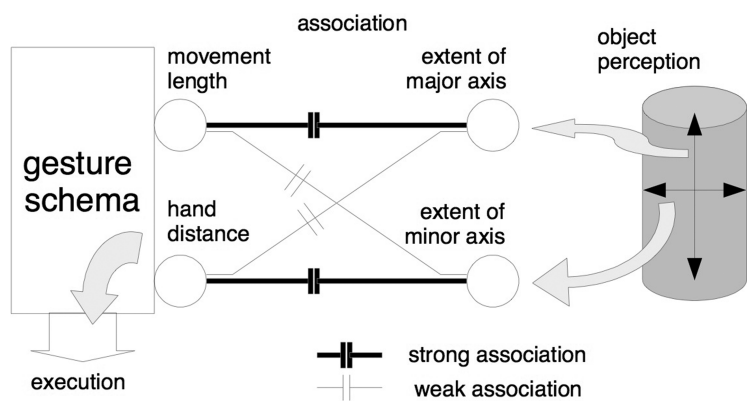


Figure 18.7 Production of gestures: object variables influence gesture form variables via learned associative links.

"growth" must include both unpacking and the focus of being through effort. The prosodic peak and the unpacking construction are not unrelated. How the language hooks the construction to prosodic peaks (supporting its presentation) is another way the GP leads to a linguistic form via a dialectic.

In contrast to the oppositions of image and linguistic categorial content in the dialectic, there is also a *synthesis* of imagery and linguistic content in the L-center. Putting this statement together with the dialectic opposition, there is both separation and unity in the generation of a cognitive state while producing an utterance, which implies two parallel layers of processes; one process to maintain the unity between image and linguistic content and another one to fuel change and development based on an opposition. In 0 and 0 we suggested using distributed representations for both visuospatial and motor imagery and to produce gestures by spreading activation that associates "active" visuospatial nodes with motor nodes via the network links. At first glance it is conceivable to apply the same concept to the representation of linguistic content. Thus, image and linguistic content could be both represented in a distributed, network-like fashion and may mutually activate each other. It is with the unity aspect of the GP that this kind of spreading activation may come into its own; unity looks native to spreading activation and might here play a straightforward role. What grows, then, is a complex, coordinated motor sequence involving the oral–laryngeal tract, breathing, and the hands/arms. This complex action is guided as it unfolds in time by "unpacking"—the construction (a template)—but its spread, focus, and peak of effort is primarily what grows from the GP.

Spreading activation seems appropriate but it will need some non-native additions. For one thing, two poles (imagery, linguistic) are needed that retain their identity and surface as a gesture and a linguistic form; one does not take over. Also, although being coexpressive, their relationship is one of opposition, not mutual activation. Their combination hence is unstable which motivates unpacking.

18.4.5 The pervasive influence of discourse context

One of the biggest challenges for a GP-compliant computational model is the influence of discourse context on the production of a gesture–speech chunk. In Section 18.2.3 we have shown how discourse context and physical factors such as hand use exert a shaping pressure on gesture. In consequence, in order to implement the shaping influence of context, the agent needs to maintain a discourse model to be able to separate out the contrastive elements(s) in a new chunk that differ from the background. Here, again, a separation of the discourse model or any other contextual knowledge source from the process of production is at odds with the theory. A partial solution that at least incorporates physical factors such as the current configuration of the hand(s) in the model is implicit in the coordinative structure approach that we suggested in Section 18.4.3. If gesture schemas or coordinative structures are bidirectional, that is if they both produce a certain class of movements and recognize them (in the training stage), then the current configuration of the body can have an impact on the activation and the selection of a schema—depending on which schema matches the current configuration best. Thus, gesture production

would implicitly depend on physical factors that do not need to be modeled and taken into consideration in a separate gesture planning stage.

18.4.6 A note on language and gesture comprehension

Though GP theory is an approach towards language production, the inverse process, forming a GP, a single unified combination of imagery and speech from verbal and gestural utterances, may occur in listeners. Speech–gesture mismatch experiments and recent neuropsychological studies (e.g. Kelly *et al.* 2004; Wu and Coulson 2005) support the assumption of a common semantic processing of gesture and speech in listeners. Still, there is a huge variety in the physical "elaboration" of gestures, suggesting that not all gestures are likely to be interpreted by a listener and that not all gestures are "designed" for the listener by the speaker. Nevertheless, there is ample evidence that some information contained in gestures reaches the addressee and contributes to the construction of meaning. Hence, cospeech gesture comprehension should be part of a complete computational model of multimodal communication with an ECA.

Gesture comprehension is already partially supported by the training-based, associative approach that we suggested in Section 18.4.3. Using the bidirectional design for gesture schemas, a schema in a listening/observing agent may respond to the gestures of a speaker and activate appropriate interpretations via the associative links. Thus, the model could be used both in a forward-chaining and in a backward-chaining manner, unifying gesture production and comprehension capabilities.

18.5 Conclusion

We examined the feasibility of operationalizing GP theory's model of language production in an embodied conversational agent. Our starting points were the theoretical and empirical underpinnings of GP theory on the one hand, and the existing computational architectures of the Max/NUMACK agents capable of producing meaningful gestures in synchrony with speech, on the other. Against this background, our analysis shows that (and how) certain aspects of non-modular GP theory can be incorporated in computational models, leading to system architectures significantly different from current approaches. In particular, we suggested a way to implement the "global" property of coverbal gestures using methods from motor control theory. Furthermore, we implied associative network models with spreading activation to implement a direct coupling between imagery and action. Such an approach could also account for the relevance of the current motor context for the selection of a gesture. What also became clear is that some core features of the theory, representation in two modes and in particular the dialectic itself, but also the inclusion of contextual factors other than the state of the motor system are currently out of reach for an explicit, detailed notion that would allow for computational modeling. It is for this reason that we believe that making efforts towards predictive computational models of a GP theoretic account of language and gesture can not only result in greatly improved conversational agents, but can also significantly further the cognitive modeling effort.

References

Barsalou LW (1999). Perceptual symbol systems. *Behavioral and Brain Sciences,* **22**, 577–660.

Bergmann K and Kopp S (2006). Verbal or visual? How information is distributed across speech and gesture in spatial dialog. In D Schlangen and R Fernández, eds. *Proceedings of Brandial 2006, the 10th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 90–7. Potsdam, September 2006. Universitätsverlag Potsdam

Billard A (2002). Imitation. In MA Arbib, ed. *Handbook of Brain Theory and Neural Networks,* pp. 556–69. Cambridge: MIT Press.

Cassell J, Bickmore T, Campbell L, Vilhjalmsson H and Yan H (2000b). Human conversation as a system framework: designing embodied conversational agents. In J Cassell *et al.*, eds. *Embodied Conversational Agents*, pp. 29–63. Cambridge, MA: MIT Press.

Cassell J, Sullivan J, Prevost S, and Churchill E (2000). *Embodied Conversational Agents.* Cambridge: MIT Press.

Cassell J, Vilhjalmsson H, and Bickmore T (2001). BEAT: The behavior expression animation toolkit. In E Fiume, ed. *SIGGRAPH 2001: Computer Graphics Proceedings*, pp. 477–486. New York: ACM Press.

Croft D and Thagard P (2002). Dynamic imagery: a computational model of motion and visual analogy. In M Lorenzo and NJ Nersessian, eds. *Model-Based Reasoning: Science, Technology, Values,* pp. 259–74. New York: Kluwer.

De Ruiter JP (2000). The production of gesture and speech. In D McNeill, ed. *Language and Gesture,* pp. 284–311. Cambridge: Cambridge University Press.

Duncan S (2002). Gesture, verb aspect, and the nature of iconic imagery in natural discourse. *Gesture,* **2**, 183–206.

Emmorey K, Tversky B, and Taylor HA (2000). Using space to describe space: Perspective in speech, sign, and gesture. *Spatial Cognition and Computation*, **2**, 157–80.

Enfield NJ (2004). On linear segmentation and combinatorics in co-speech gesture: A symmetry-dominance construction in lao fish trap descriptions. *Semiotica,* **149**, 57–123.

Firbas J (1971). On the concept of communicative dynamism in the theory of functional sentence perspective. *Philologica Pragensia,* **8**, 135–44.

Glasgow JI (1993). The imagery debate revisited: a computational perspective. *Computational Intelligence,* **9**, 309–33.

Hadar U and Butterworth B (1997). Iconic gestures, imagery, and word retrieval in speech. *Semiotica*, **115**, 147–72.

Harris R (2002). Why words really do not stay still. *Times Literary Supplement* 26 July, 30.

Kelly S, Kravitz C, and Hopkins M (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, **89**, 253–60.

Kelso J, Tuller B, and Harris K (1983). A "dynamic pattern" perspective on the control and coordination of movement. In P McNeilage, ed. *Speech Production,* pp. 137–73. New York: Springer.

Kendon A (1980). Gesticulation and speech: Two aspects of the process of utterance. In M Key, ed. *The Relationship of Verbal and Nonverbal Communication,* pp. 207–27. The Hague: Mouton.

Kopp S and Graeser O (2006). Imitation learning and response facilitation in embodied agents. In J Gratch *et al.*, eds. *Intelligent Virtual Agents 2006,* pp. 28–41. Berlin: Springer-Verlag (LNAI 4133).

Kopp S, Tepper P, and Cassell J (2004). Towards integrated microplanning of language and iconic gesture for multimodal output. In R Sharma and T Darrell. *Proceedings of the International Conference on Multimodal Interfaces (ICMI'04)*, pp. 97–104. New York: ACM Press.

Kopp S and Wachsmuth I (2004). Synthesizing multimodal utterances for conversational agents. *Journal of Computer Animation and Virtual Worlds*, **15**, 39–52.

Kosslyn SM (1980). *Image and Mind.* Cambridge: Harvard University Press.

Kosslyn SM (1987). The medium and the message in mental imagery—a theory. In N Block, ed. *Imagery,* pp. 207–44. Cambridge: MIT Press.

Leβmann N, Kopp S, and Wachsmuth I (2006). Situated interaction with a virtual human—perception, action, and cognition. In G Rickheit and I Wachsmuth, eds. *Situated Communication*, pp. 287–323. Berlin: Mouton de Gruyter.

Levelt W (1989). *Speaking.* Cambridge: MIT Press.

McNeill D (1992). *Hand and Mind: What gestures reveal about thought.* Chicago: University of Chicago Press.

McNeill D (2005). *Gesture and Thought*. Chicago: University of Chicago Press.

McNeill D and Duncan SD (2000). Growth points in thinking-for-speaking. In D McNeill, ed. *Language and Gesture,* pp. 141–61. Cambridge: Cambridge University Press.

Müller C (1998). *Redebegleitende Gesten. Kulturgeschichte—Theorie—Sprachvergleich*. Berlin: Berlin Verlag.

Nobe S (2000). Where do most spontaneous representational gestures actually occur with respect to speech? In D McNeill, ed. *Language and Gesture*. Cambridge: Cambridge University Press.

Reiter E and Dale R (2000). *Building Natural Language Generation Systems*. Cambridge: Cambridge University Press.

Saltzman E and Kelso S (1987). Skilled actions: a task-dynamic approach. *Psychological Review,* **94**, 84–106.

Saussure F (2002). *Écrits de Linguistique Général* (compiled and edited by S Bouquet and R Engler). Paris: Gallimard.

Slobin D (1987). Thinking for speaking. In J Aske, N Beery, L Michaelis and H Filip, eds. *Proceedings of the Thirteenth Annual Meeting of the Berkeley Lingusitic Society,* pp. 435–445. Berkeley: Berkeley Linguistic Society.

Sowa T (2006). Towards the integration of shape-related information in 3-D gestures and speech. In F Quek and J Yang. *Proceedings of the Eighth International Conference on Multimodal Interfaces,* pp. 92–9. New York: ACM Press.

Sowa T and Wachsmuth I (2005). A model for the representation and processing of shape in coverbal iconic gestures. In K Opwis and IK Penner, eds. *Proceedings of KogWis05. The German Cognitive Science Conference 2005*. Basel, Switzerland. Schwabe Verlag.

Streeck J (unpublished). Gesture: The manufacture of understanding.

Vygotsky LS (1987). *Thought and Language*. Edited and translated by E Hanfmann and G Vakar (revised and edited by A Kozulin). Cambridge: MIT Press.

Wu YC and Coulson S (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology*, **42**, 654–67.