

Modeling Dialectic

David McNeill
University of Chicago

Abstract abstract

- Gestures synchronize with semantically and pragmatically coexpressive linguistic segments.
- The question is how speech-gesture synchrony is achieved.
- In the growth point, speech-gesture synchrony is part of thinking itself and arises from an imagery-language dialectic.
- This dialectic is inherently dynamic, unstable, and seeks change, which is its psycholinguistic function.
- How can a dialectic be modeled? It cannot be captured with imagery descriptive features, which lose the opposition of modes essential to a dialectic. The result, although a system that ‘behaves’, is static and purely structural.
- A solution is to treat imagery as a kind of action, which can capture the global property of imagery but has limitations of its own.

- My approach is via behavioral science.
- Based on the close observation of behavior as a multimodal stream, including speech, gesture, posture, and social interaction
- Here, specifically, I will focus on speech and gesture.

Purposes

- I'll start with 'Max' - an animated agent with gesture capabilities inspired by Ipke Wachsmuth at U. Bielefeld
- My comments are based on extensive discussion with the developers, students of Wachsmuth, Stefan Kopp (production side) and Timo Sowa (comprehension side), and my gesture colleague, Sue Duncan.

Very different agendas

- Formatting a theory of human language production on the basis or close systematic observation of natural multimodal language behavior
- Programming a virtual human to speak and gesture in ways that human observers will feel is ‘natural’

Max

- Began as testbed for modeling speech and gesture
- Now used in various settings
 - Museum guide
 - Collaborator in VR assembly. Clips illustrate the latter:



Max's generation pipeline

Action selection and content planning

Selecting acts to perform and organizing them into a structural plan, choosing domain-specific knowledge

Behavior planning

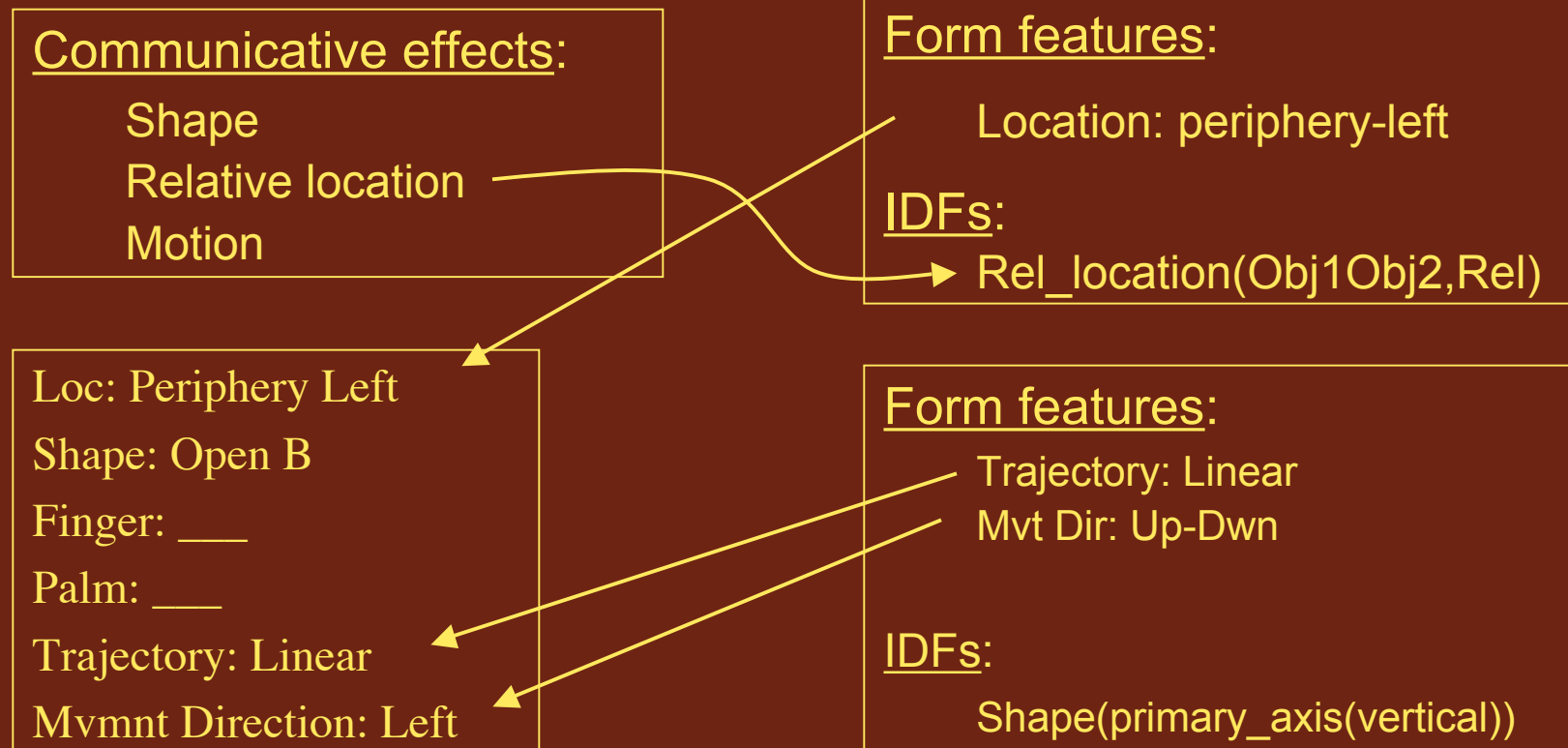
Taking the plan and recoding each step (act) into surface form of coordinated multimodal behaviors that realize it

Behavior realization

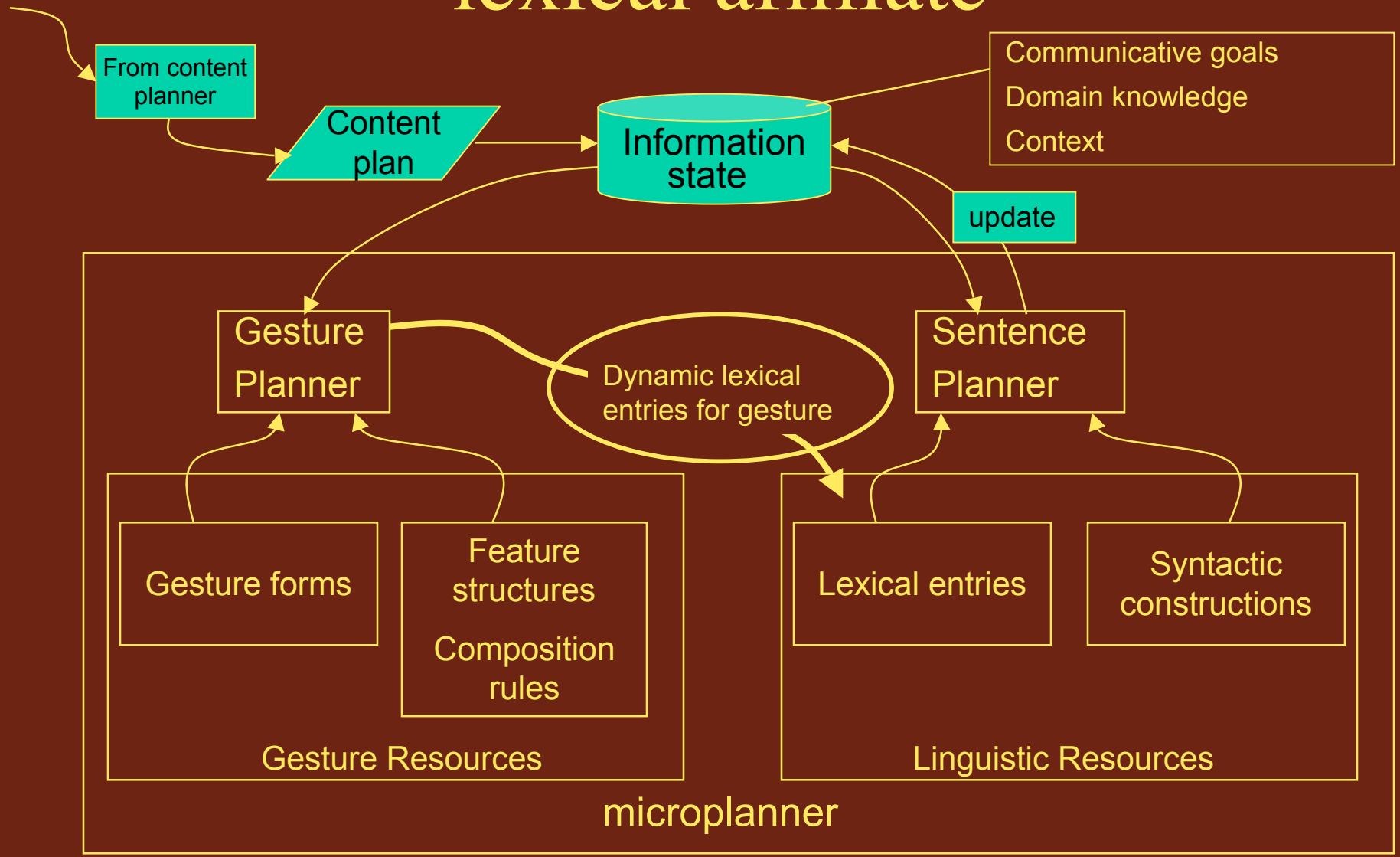
Turning linguistic structures and nonverbal behavior into synthetic speech and animation of face and body

Gesture planner

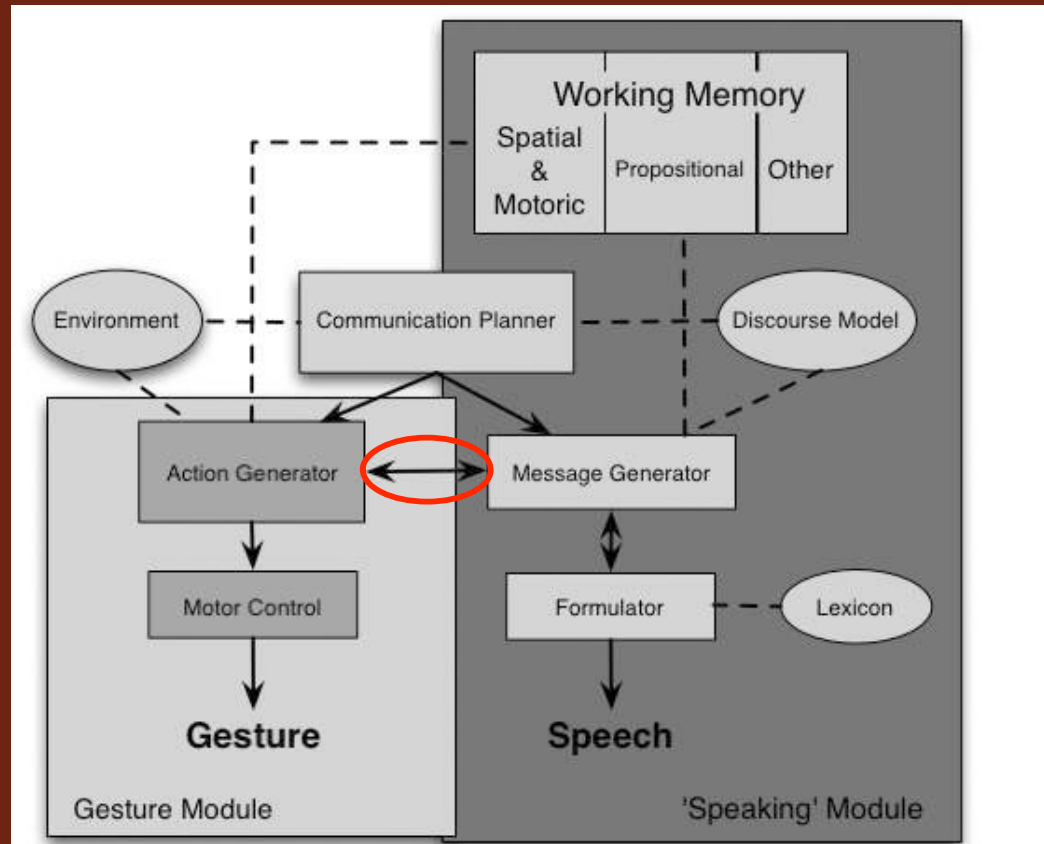
- Builds a gesture feature structure by selecting Form Feature Entries whose Image Description Features match the desired communicative effects



One solution to timing - gesture picks lexical affiliate



Another solution - speech and gesture separate and exchange signals



Based on Kita and Özürek (2003), Figure 7. Dashed lines are information sources, solid arrows are inputs. The darker box on the right is Kita and Özürek's version of Levelt's 'Speaking' model (1989), while the lighter box on the left is their added 'gesture module.'

Stroke-speech timing: a good test case

- Max works as follows – looks ahead, sees what the lexical affiliate will be, calculates how far back the preparation will have to be in order for the stroke to coincide with the lex aff. Then speech and gesture are generated on their own tracks, and the two kept in synch by cross-signals. This is basically static: Max gives pathways for change, but requires a stimulus to get going. The Kita-Özyürek model presumably does something similar.
- In contrast, in the GP the gesture image and linguistic categorization constitute an idea unit, and timing is inherent in constituting this thought. The start of prep is the dawn of the idea unit, which is kept intact and unpacked, as a unit, into a full utterance. Change is inherent to the unit.

The Growth Point Hypothesis

- Radically different concept of synchrony in the GP
 - The synchrony of gesture and language is *part of the idea unit*, it is inherent to the thought.
- This is because the GP is a unit formed out of two opposed cognitive modes - imagery and linguistic categorial content for the same idea unit.
- To have an idea in this model, the meaning exists in two semiotic modes at once. Thus, image and co-expressive linguistic content must be synchronized: gesture-speech synchrony is inherent to thought.

The GP

- The GP is an empirical as well as theoretical concept
- Growth points are inferred from the totality of communicative events with special focus on speech-gesture synchrony and co-expressivity.
- Called ‘growth point’ because it is meant to be the idea unit starting point of utterance formation.

Definition of imagery

- Don't think of a photograph or film strip
 - An image is actional as well as visuo-spatial.
- The defining qualities are:
 - Global-synthetic, to be explained later.
 - Non-combinatoric (the parts relate to one another but do not combine hierarchically).
 - No standards of form.
 - Instead, *form is determined by meaning.*
- So the ultimate def is: an image is meaning embodied in form. There is no separate standard of form.



Image of swinging on a rope

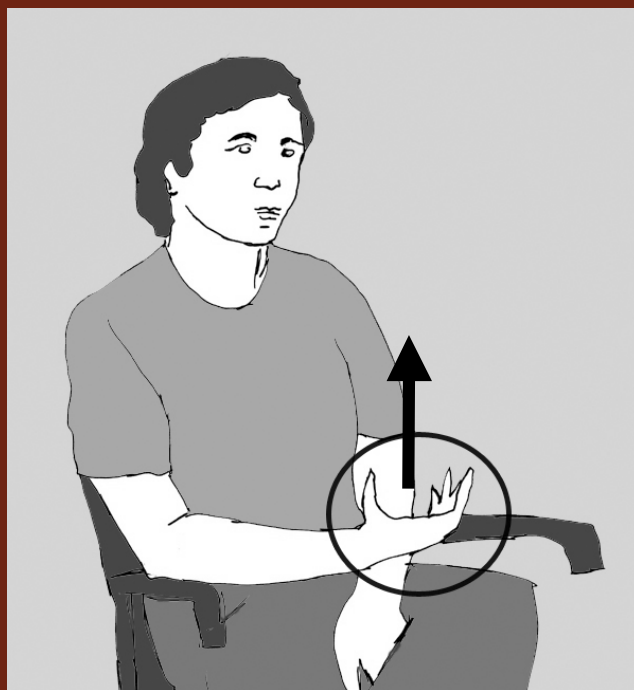
Why ‘growth point’ - the name?

- Meant to be the initial form of thinking for (and *while*) speaking, out of which a dynamic process of organization emerges.
- A theoretical unit in which the principles that apply to the mental growth of children—differentiation, internalization, dialectic, and reorganization—also apply to real-time utterance generation (in both adults and children).
- The concept that there is a specific starting point for a unitary thought. Although an idea unit may emerge out of the preceding context and have ramifications in later speech, it does not exist at all times. It comes into being at some specific moment; the growth point is this moment, theoretically.

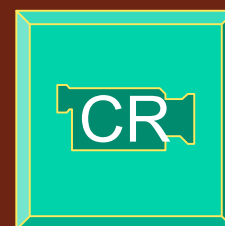


A growth point example

The GP is something like “upward moving hollowness”, categorized linguistically as “up through”.



Our method:
narration of a
known stimulus



The GP is an image with a foot in the door of language; equally, it is language embodied in an image. Even when the meanings seem close, they are in opposite modes of cognition.


Imagery-language dialectic

- The simultaneous activation of imagery and linguistic encoding creates a dialectic.
- A dialectic is a combination of opposites that fuels thought and speech. Without it, thought and speech require an external stimulus, the condition in an experiment but is not usually in discourse.
- Instability is an essential feature of thought according to the dialectic - a key to the dynamic dimension of language - and for this reason speech-gesture synchrony is *inherent to thought*.


Minimal Unit Of Dialectic

- The Growth Point is the smallest particle of dialectic. In Vygotsky a ‘minimal unit’ is contrasted to an ‘element’.
 - A GP is a ‘minimal unit’ in that it retains the property of a whole (thus not reduced to an ‘element’) Min units are the smallest components that retain the essential property of being a whole.
 - An ‘element’ is the more familiar product of reduction to simpler components, without the requirement that it be a whole - in fact, requiring that it not be a whole but a lesser component.

The opposition

- The unlike cognitive modes at the same time:
 - Imagery = global and synthetic. 
 - Linguistic encoding = combinatoric and analytic.
- In the example
 - Imagery: The hand = Sylvester, the open shape = interiority, the direction = up the pipe, motion = ascent. These values are globally derived - not form features with pre-established meanings.
 - Linguistic encoding: words, “goes”, “up” and “through” are meaningful on their own, and by combining, build up the meaning of the whole. “This time” metapragmatically encodes the contrast of the second ascent to the first

Global and synthetic

- *Global* = the determination of meaning proceeds downward. The meanings of the ‘parts’ of the gesture are determined by the meaning of the whole. In fact, parts have no reality except in the meaning landscape of the whole. So, for example,
 - the hand = Sylvester, but the individual fingers mean nothing;
 - or in a different case, the two hands mean ‘hands’ – the parts depend in both cases on the global significances of their gestures).
- This semiotic model contrasts to the upward determination of meanings in sentences, which requires an independent listable, recurrent morphology and syntax.
- *Synthetic* refers to the fact that a single gesticulation concentrates distinct meanings that spread across the surface of the accompanying speech.

Earlier - Wundt and Saussure (1910s)

- Two earlier figures who saw the duality of language and cognition: Wundt & Saussure
 - Wilhelm Wundt - the ‘father of experimental psychology’
 - Ferdinand de Saussure ‘the father of modern linguistics’
- The idea an essential duality is a view of language that has been lost for nearly a century

Imagery is dynamic - shaped to fit significance

- Speakers who omit S's first ascent, on the outside, mentioning only the second, on the inside, do not shape the gesture to convey interiority. For such a speaker, interiority isn't a point of contrast, even though it was part of the perceived stimulus. And it does not come into the thought unit.

context & psy pred



1. speakers (two) remember inside only => no interiority. (Speaker 2: thumb only.)
2. speakers (two others) who remember both => interiority

Growth point contexts

- A further source of dynamic change - the growth point is inseparable from its context. Context is inherently non-categorical. In the example, the field of potential oppositions was How to Get Up the Pipe.
- This was Sylvester's second try using the pipe to get to Tweety. The context at the second ascent included the speaker's description of the first ascent
 - The point of differentiation highlighted the interiority of the pipe, whereas the first 'climbs up' showed the pipe from the outside.
 - In the 'rising hollowness' example shown at the start, when the speaker embodied the concept of ascent and interiority in a single image, the point of differentiation was again the inside.
- The gestures in both cases plus their linguistic categorizations highlight precisely this differentiating factor, which is the core idea unit at the moment, in the context.

How the GP is Formed - Differentiation

- To understand the GP we need to analyze the background or context against which this differentiation occurs.
- The GP is formed by differentiation from the context. Rather than start as a unit and then fit into context, a GP depends on a context to start.
- The GP is the point of contextual weight and newsworthiness, the significant departure of content in the immediate context of speaking.

The Psychological Predicate - Key to explaining differentiation and context

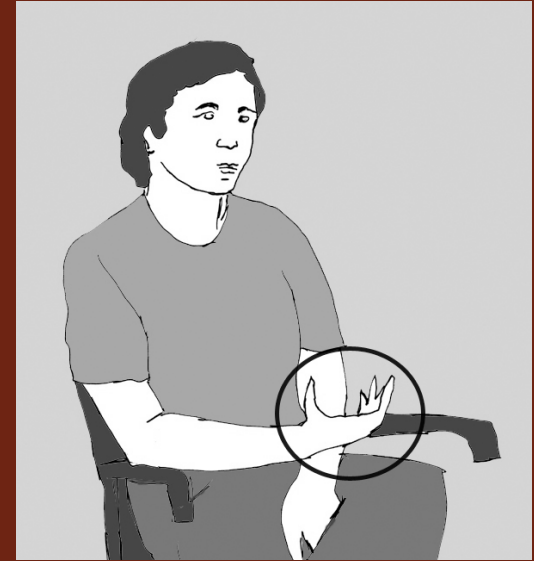
- Psychological predicate - not necessarily a grammatical predicate.
- Vygotsky: Marks a significant departure from the immediate context.
 - ◆ “What happened to the clock?” - “It fell”
 - ◆ “What fell? - “The clock”
- Implies context as background.
- We see differentiation when the psych pred (GP) makes something stand out from a background - inside vs. the previous outside attempt in the #1 cases, no such contrast in the #2 cases. [back to mov](#)

The global property & features

- The global character of the gesture is part of how it stands out from its context.
- Because the gesture differentiates a point of contrast as a whole, a feature (interiority) emerges as significant - the feature is non-existent without the whole.
- The first step is not the feature but the global significance.
- Thus forms can turn out to have quite different values, depending on context.

For example

- The same upward, fingers spread hand shape has a completely different significance in another gesture.
- Here, a similar hand shape is a metaphor for the 'basic plot line' - the hand is a surface (not an interior space), and upwardness is supporting the idea of the plot line (not ascent)
- In other words, the significance of the gesture 'parts' - what even counts as a part - derives from the significance of the whole.
- Both gestures are points of differentiation, but the context and hence significance decides what 'parts' like palm up signify.



VJ met

The problem

- The problem is that the use of features in computational models appears to force the process of gesture creation to be combinatoric, thus losing the opposition of semiotic modes essential to the dialectic - in fact, reducing the GP to Vygotskian ‘elements’.
- To be global, the process wants to work from the overall meaning downward.
- Even if we force a model to proceed in this direction, form features need to have their own meanings in order for the model to find them – but do they?

- Other modeling approaches have the same limitation.
 - Barsalou's perceptual symbols, for example - have have the goal of giving imagery the same semiotic as language. This makes a dialectic impossible to model.
 - Connectionism: reduces conflict to a single mode
 - Spreading activation: similarly, reduces to a single mode
- All have the goal of achieving computational adequacy, but do not have the capability of modeling the dialectic

Lex affs

Is action the solution?

- Suppose that a speaker improvises something that we, on analysis, decide means ‘interior’, ‘upward’, and ‘effort’ – what does she need to do for this?
- She needs to perform *an action that embodies these meanings*. Does this imply combining form-meaning features? *Or is it enough to ‘act’?* Is the action of rising upward inside the pipe sufficient to generate a gesture with the significances we are after?
- This would be an image in the sense of being meaning-determined and global-synthetic.

So, the resolution

- The idea of *coordinative structures* seems to apply, with the addition of a thought-language-hand link (accessing and steering coordinative structures using significances).
 - Ideas or significances are *attractors* of coordinative structures; the coordinative structures zero in on these attractors. Meanings coordinate actions to make gestures (whereas regular actions are coordinated by goals).
 - The existence of a ‘thought-language-hand link’ in the human brain is suggested by the remarkable case of IW, a man suddenly deafferented from the neck down, who still makes gestures but not instrumental actions. IW
- The properties of the attractor bring out features 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.
- This role for coordinative structures under the spell of significance also is compatible with the point that speech-gesture synchrony is inherent to thought.

- Coordinative structures are not themselves significant forms; they are “flexible patterns of cooperation among a set of articulators to accomplish some functional goal” (anonymous Yale web handout).
- There is no lexicon of feature-meaning pairs (‘facing down → force downward’ and the like). 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 are the outcomes of the gesture, not the source.
- Each coordinative structure is an ‘action primitive’, but the critical difference from a feature is that coordinative structures do not have significances - they are just bits of action lying about and ready to use. limits

Limits

- A weakness of the coordinative structures approach is that it implies a distinction between ‘image’ and ‘gesture’ (the attractor is the image and coordinative structures fashion a gesture to embody it).
- I think this distinction is wrong; the gesture *is* the image, not a copy of it – it is the image in its most material form.

This seems to be a limit on our scientific language

- ‘Coordinative structures’ and ‘affordances’ imply a distinction where none is (affordances lure coordinative structures).
- We need a different way of speaking, and possibly Maurice Merleau-Ponty, the philosopher of phenomenology, provides it; some suggestive remarks:

So, here is a try

- It's not that there is an image-gesture distinction; *significance is a way of orchestrating coordinative structures*, and *this* is what we call an 'image'.
- Thus, as Merleau-Ponty says, "speech is thought" and features like space and one's body are not elements:
 - "We must recognize first of all that thought, in the speaking subject, is not a representation, that is, that it does not expressly posit objects or relations. The orator does not think before speaking, nor even while speaking; his speech is his thought."
 - "I do not need to visualize external space and my own body in order to move one within the other. It is enough that they exist for me, and that they form a certain field of action spread around me. In the same way I do not need to visualize the word in order to know and pronounce it. It is enough that I possess its articulatory and acoustic style as one of the modulations, one of the possible uses of my body."

But still not a model

- Coordinative structures explain the global property, essential to the dialectic but not:
 - The growth point itself;
 - The differentiation of psychological predicates;
 - Growth;
 - Inseparability from context;
 - Co-presence of imagery and linguistic categorization;
 - The coexpressiveness of imagery and language;
 - Internal tension and motivation;
 - Change/unpacking, although aspects of max apply to unpacking.

Why should CS be interested?

Two broad approaches:

1. Create a behavior stream that uses technical fundamentals – a cognition engine, a motor engine, a perceptual engine, etc., but does not consider a specific theory (though one usually is implied)
2. Take a behavioral process as the goal and see whether computational models are able to explain it. If not, what are the obstacles? This can be a challenge for CS innovations.

I have been illustrating the second approach. Clearly it is of interest to the gesture linguist. Discussions of imagery in the AI literature, so far as I am aware, invariably posit some lexicon of perceptual features, and this is the challenge: how to model global-synthetic imagery.

New information systems?

- I don't know that coordinative structures have been modeled, so can't see if they avoid the use of features, but to use action schemes as a route to modeling thought, is one implication of the gesture work.
- Conceiving of representation as a kind of action capable of orchestrating coordinative structures based on meaning, may open the way to model the global property and, by combining it with a symbolic representation of the same information, model the dialectic.
- Also, define co-expressiveness across semiotic modes. A hybrid analog-digital machine? in case
- A further aim: create self-defining, self-segregating imagery solutions that may model growth points and how they form out of contexts.
- Are these steps feasible? This a CS question but it is essential for modeling human language (as opposed to programming a virtual human).

Desired elements in an analog device

- These points can provide something like imagery in an analog device that could engage in a dialectic.
 1. 3D
 2. Orientation
 3. Direction
 4. Texture
 5. Spatial array
 6. Local identity (granularity okay)
 7. Memory
 8. Organized as action (perhaps coordinative structures)



Many Thanks - The End!



Vygotsky c. 1930