Gesturing Through Time: Holds and Intermodal Timing in the Stream of Speech

by

Mischa Alan Park-Doob

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Linguistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in Charge:

Professor Eve E. Sweetser, Co-chair
Professor William F. Hanks, Co-chair
Professor Keith Johnson
Professor Carla L. Hudson Kam

Fall 2010

Gesturing Through Time: Holds and Intermodal Timing in the Stream of Speech

by

Mischa Alan Park-Doob

Abstract

Gesturing Through Time: Holds and Intermodal Timing in the Stream of Speech

by

Mischa Alan Park-Doob

Doctor of Philosophy in Linguistics

University of California, Berkeley

Professor Eve E. Sweetser, Co-chair
Professor William F. Hanks, Co-chair

Most previous work examining co-speech gestures (the spontaneous bodily movements and configurations we engage in during speaking) has emphasized the importance of their most salient or energetically expressive moments, known as gesture ‘strokes’ (Kendon 1980). In contrast, in this dissertation I explore the potential functions of intervals of gestural *stasis,* or gesture ‘holds’, in which the hands or body maintain particular configurations across variable spans of time, interwoven with the stream of speech. Through the embodiment of a constant form within continuously evolving face-to-face interactions, holds make possible a unique and understudied array of functions relating to the maintenance of ideas and contexts across time.

Chapter 1 introduces the corpus of videotaped dyadic conversations from which all of the examples are drawn, discusses the history of the concepts of ‘stroke’ and ‘hold’, and illustrates the structural possibilities for the timing of holds with respect to co-expressive speech: they bear content that is not just simultaneous with, but also ‘retrospective’ and/or ‘prospective’ of, portions of the full composite utterances in which they occur.

Chapter 2 illustrates that holds lasting across pauses and disfluencies support continued expressiveness and interpretability, alternately presaging new content that will also be part of a fluent resumption, or maintaining retrospective links to prior content that can contextualize the resumption.

Chapter 3 discusses the frequent expressive complementarity of co-timed speech and gesture, as it relates to the debate on speech-gesture synchrony, and further demonstrates that preliminary commitments to utterances are often partially *fulfilled* from the earliest

moments because of gestural cues that are interpretable at all points of their lifecycles, including preparatory phases.

Chapter 4 discusses the implications for attention and memory of gesture holds acting as temporary cognitive artifacts, forming ‘bridges’ across interruptions and competing representations by interlocutors, thereby functioning retrospectively as ‘recall cues’ to previous moments of the interaction.

Chapter 5 focuses on instances of gesture holds combined with listener-directed gaze that are maintained across turn transitions, then released, allowing speakers to ‘hand off’ control while enforcing a context for the next turn.

Chapter 6 synthesizes the preceding chapters and suggests directions for future research.

Table of Contents

List of Figures

Index of Examples

Key to symbols and annotations

Gesture annotation conventions

(RH) “Right Hand”: When necessary for clarity, this marker is placed to the left of a line of gesture annotation representing the gesturer’s right hand.

(LH) “Left Hand”: When necessary for clarity, this marker is placed to the left of a line of gesture annotation representing the gesturer’s left hand.

(BH) “Both Hands”: When necessary for clarity, this marker is placed to the left of a line of gesture annotation representing both of the gesturer’s hands moving in unison.

(A’s) Participant A’s gestures.

(B’s) Participant B’s gestures.

_____ Underscore: This marking spans intervals in which the hands or arms are at rest.

↗ Marks the onsets of preparation phases (departures from rest, or departures from previous gestures).

^ Marks the apex of major gesture movements (gesture movement ‘strokes’).

ˆ Marks small, pulse-like excursions (gesture ‘beats’).

‾‾‾‾‾‾ Superscore: This marking spans intervals in which a gesture configuration is held in place (gesture ‘holds’).

┄┄┄┄┄
‾‾‾‾‾‾ Dashed line above superscore: Indicates intervals of ‘waggling’ occurring during gesture holds (emphatic, continuous small movements, often consisting of small back-and-forth or twisting motions of the wrists).

‾‾‾‾‾‾ Gray color applied to an existing marking indicates a ‘sagging’ gesture configuration, usually during holds, in which muscle tension fades, arms lower slightly, and hands loosen.

↘ Marks the onsets of retraction phases (movements toward rest position).

Speech transcription conventions

A: Placed to the left of the transcript of Participant A's speech.

B: Placed to the left of the transcript of Participant B's speech.

⎡**A**: ⎤ Brackets that span two lines indicate simultaneous speech by both partici-
⎣**B**: ⎦ pants. Unbracketed lines occur sequentially.

· ·· ··· Raised "pause dots" mark silent pauses and hesitations occurring in the flow of speech. Not intended to be precise, each dot can be taken to represent roughly 0.1-0.2 seconds, with more precise figures given as needed.

n·· Pause dots directly adjacent to a segment indicate a pause which is partly filled by a continuation of that speech sound.

# Marks a pause accompanied by an audible intake of breath.

* Marks moments of disfluently aborted speech.

@@ Marks periods of laughter.

(...) Indicates that a portion in the middle of the transcript has been omitted due to lack of sufficient horizontal space.

{*swallow*} Descriptions in curly brackets represent other time-filling audible or visible events involving the vocal tract.

, . ? ! Standard punctuation is included to improve ease of reading. These four marks can roughly be taken to indicate, respectively, non-final intonation, sentence-final intonation, high rising intonation, and exclamatory speech, in addition to providing very rough hints about speaking rate.

General architecture of all examples

Top: The full speech transcript alone, often including material that precedes the moments coinciding with the images, to help contextualize them.

Middle: A prose description of the gestural behavior occurring in the example.

Bottom: A sequential set of still images, lined up with a limited reprinting of the speech transcript, in addition to gesture annotation. The images are best examined in conjunction with the preceding prose description.

Annotations marking simultaneity of elements in the unfolding scene

Moments of transcript corresponding to a presented image are **shaded** and have an alphanumeric index placed above them. For example:

c

the · train will come in

The same index is displayed to the left of the section of prose describing the gestural behavior of which the image is a snapshot. For example:

c: *BH have now finished the preparation phase for a BH forward-thrusting stroke which will time with the end of the word "will."*

The same index is also displayed above the image in question, aligned vertically with the shaded section of reprinted transcript to which the image corresponds:

c

the · train will

The indexes in each example proceed sequentially: "a, b, c, d" and so on.

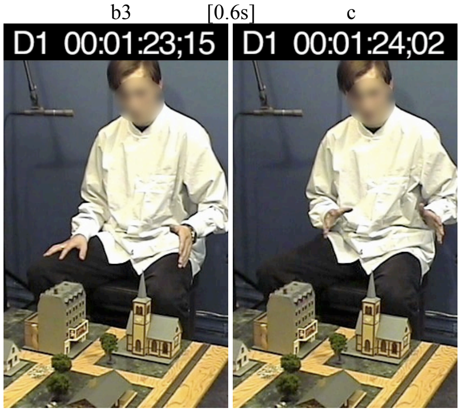
Often, an index letter will be repeated with appended numerical suffixes in order to indicate that a subset of the images form a coherent grouping. For example, a set of indexes [a b1 b2 b3 c] would be appended to a set of five sequential images, with the middle three images forming a coherent subgroup.

Furthermore, the prose description of gestural behavior will often refer to a range of snapshots, rather than just a single image. For example:

b1-b3: *LH stays almost in the same position through the disfluency, but gradually rises an additional inch as the palm slowly rotates to vertical, solidifying into a more clearly defined deictic even as he fails to restart fluent speech.*

Time annotations

The amount of time separating any two sequential images is given to the nearest tenth of a second, in square brackets, above and between the images:



The timecode stamped on each image shows the Session Number (a unique identifier of each separate conversation in the corpus), followed by standard NTSC timecode in [hours : minutes : seconds ; frames] as measured from the starting ‘clap’ of each file in the corpus (each session was initiated by the investigator wielding a movie-style “clapper” while announcing the date and session number).

A note on NTSC ‘drop-frame’ timecode:

NTSC video (the nearly-obsolete broadcast standard for standard definition video in the U.S. and Japan) runs at a real speed of 29.97 frames per second, but displays with 30 frames devoted to each nominal second of timecode. Because the video actually runs slightly slower than 30 frames per second, each second of timecode actually lasts 1.001 seconds. The timecode therefore utilizes the “drop-frame” system of counting: the first two frame *numbers* of each whole minute are skipped over (except when the minute is divisible by 10), and this keeps the timecode display in line with real elapsed time. The “drop-frame” convention is indicated by the use of a semicolon, instead of a colon, before the frame number. For example, the next frame of video occurring after [00:00:59;29] will be tagged with a timecode of [00:01:00;02], and there will not be any frames of video labeled [00:01:00;00] or [00:01:00;01] (note that no frames of video *content* are actually “dropped”). These technicalities can be safely ignored by the reader: when making rough calculations of elapsed time between different video stills in the same short snippet of conversation, it is safe to assume that three frames of video correspond to a tenth of a second (100 milliseconds), and the question of skipped frame numbers can be ignored. Any resulting error will be too small to be of consequence.

A note on the origins of the symbols used for transcription and annotation:

The scheme for gesture annotation, as well as the general layout of the examples and their alignments for marking simultaneity, are entirely of my own invention. The speech transcription conventions, however, are a mixture of practices employed by the McNeill Lab Center for Gesture and Speech Research at the University of Chicago, and conventions which I adapted from Chafe (1994) and which are part of the system of discourse transcription developed by Du Bois and colleagues in the late 1980s (see Du Bois et al. 1993). McNeill Lab conventions include the use of "#" for audible inhalations and "*" to mark speech trouble. Inherited from Du Bois et al. are the use of "@" to mark laughter, and small dots to mark pauses.

Acknowledgments

I would never have predicted the winding path leading me to take on this project, but I'm glad it's all worked out the way it has. Looking back at what first drew me into the world of gesture studies, it was something as simple as this: I was an undergrad, enjoying a bite to eat at the University of Chicago linguistics department's famous weekly "tea," and graduate student Chris Corcoran must have heard me making some remark about how it would be nice to have a summer job. I should therefore begin by thanking her for introducing me to Karl-Erik McCullough and David McNeill, who let me loose amid the multiple cameras, blue-painted walls, and assorted Sylvester and Tweety paraphernalia that are a staple of any project taking place in the McNeill Lab. Its unique character and intellectually welcoming atmosphere, even to a youngster like I was at the time, have been a source of continuing inspiration in the years since.

Although I soon left for graduate studies at UC Berkeley, many McNeill Lab members have influenced my thinking, and set me on the successful path that led me to this point, including Karl-Erik McCullough, Susan Duncan, Fey Parrill, Nobuhiro Furuyama, Amy Franklin, Irene Kimbara, Mika Ishino, Gale Stam, and I'm sure several others who I will later be mortified to have forgotten to name. I also thank Francis Quek for setting up the NSF/KDI project at the University of Chicago and Wright State University/Virginia Tech, whose grant provided the funds for the project that led to my employment, as well as the videos from which I have drawn all my examples.

In more recent times I have benefited tremendously from David McNeill's willingness to discuss my research with me, in person and over email, especially his patience with my occasionally wrong-headed readings of his own theories, and I am also particularly grateful to Susan Duncan for her incisive comments on several iterations of the approach I have taken. Most of all, I thank Karl-Erik McCullough for providing me with much of the intellectual inspiration behind this work in his own dissertation from five years ago, and for all the great discussions we have had before and since, not to mention his willingness to sit down and help me polish up a few things in these last few days, here halfway round the world from where it all began.

In Berkeley, I feel tremendously lucky to have had a dissertation committee willing to afford me so much freedom, allowing me to take this project in what is without a doubt an unusual direction compared to most work one finds in a linguistics department, even one as unrestrictive as Berkeley's. Keith, Carla, Bill, and Eve: I've always admired your work, and it's been a pleasure working with you over the years. I'm especially grateful to Eve Sweetser for her sure and steady advocacy of gesture studies and cognitive linguistics at Berkeley, both as my co-chair and as leader of the Berkeley Gesture Group, whose members have all contributed to my thinking over the years. Although I will not try to name everyone in that group who I've had the privilege of bouncing my ideas off of, I would like to single out Nathaniel Smith for his early work on persistent structure in gesture, which has informed my approach, and Katharine Young for so many fascinating and encouraging discussions on the phenomenology of multimodal interaction, besides several others I name below. More recently my co-chair Eve Sweetser has also helped tremendously during these final stages of work as I have put the final pieces into place, on an accelerated schedule that I was lucky to find my committee willing to follow. Belén Flores, meanwhile, is in many ways the true force behind every successful graduate student in Berkeley's linguistics department, much more than her humble staff title would indicate. I'm sure none of us has any idea how we would have gotten to the end if not for her eternally cheerful, seemingly effortless guidance and competence through every step.

Turning to my adopted 2nd department, I'm also especially grateful to my co-chair Bill Hanks, not just for his close interest and detailed commentary on this dissertation, but for inviting me to join him in the most intellectually stimulating research assistantship I could have asked for, and many hours worth of helpful mentorship over the years, full of encouraging and profound discussions on my research as well as his. I'll think back fondly on these years, that's for sure. Among his sphere of students and postdocs in linguistic anthropology, I've enjoyed my close contact in recent times with Shaylih Muehlmann, Terra Edwards, Mara Green, Nathaniel Dumas, Xochitl Marsilli, and others I'm sure I'm neglecting, with whom I've grown to feel that I was engaged in something exciting and worthwhile, and learned not to be afraid of heading in directions far afield from where my departmental affiliations may have started me.

Outside of my home departments, I'm also very grateful to have met Nick Enfield and had the chance to discuss my work with him and absorb his ideas, so many of which bear directly on my efforts here.

Finally, thanks to all my other colleagues in the linguistics department at Berkeley, especially Elisabeth Wehling and Elise Stickles of the Gesture Group, confederates in this crazy enterprise we pursue, Alex Bratkievich, kindred soul and fellow film lover, and housemates Karen Sullivan, Eric Prendergast, and Roslyn Burns, who've all contributed in turn to the uniquely homey atmosphere of our little cottage in a garden full of cats and fruit trees. Karen and Eric, I'm glad we had the chance to live by each other for a while; it's been fun and I hope you stop by to watch a film sometime. And Roslyn, thanks for all the food donations and good humor these last few weeks, and sorry about gobbling up all the chorizo the other day… it was there and I was ravenous. To Berkeley's many coffee shops, I hope you make good use of all those thousands of dollars – I know I certainly made good use of your tables, outlets, and caffeinated beverages. And thanks of course to my parents, Deborah Doob, Robert Park, and Barbara Park, and my siblings Robin Park-Doob, Ian Park, and Casey Steinau, and their endless patience and encouragement of all my endeavors. I'm lucky to have you guys around.

Frankfurt an der Oder
July 27, 2010

1 Introduction

Each of us lives in a world of inescapable temporal and physical constraints. The flesh of our bodies is real and bounded, and the passage of time is inexorable, with every facet of observable reality in some stage of both emergence and decay. How do these constraints manifest themselves in our lives as social beings, and what sorts of adaptations might we have developed in our daily behaviors and strategies? How do the capacities and limitations of our attention and memory interact with the semi-permanence of our modifiable environment? I hope to provide a few small answers to these questions in the chapters to come, since it became clear during my investigation that much of what I was seeing might be a consequence of universal or near-universal facts of human physiology and cognition, coupled with independent physical and temporal truths about the reality we inhabit.

There is a real need to highlight some of these immutable features of reality, to avoid recasting them as human traits. Enfield and Levinson (2006) have endeavored to synthesize the major findings of sociology, anthropology, linguistics, and psychology into a new science of "human sociality," one which identifies the common ground in our individual cognitive capacities, our systems of interaction, and our cultures. A relatively limited set of building blocks is proposed, from which all our social practices are suggested to have emerged, with a major theme being the universality across cultures of many basic mechanisms. They propose that "culture can reach deep down into the details of interaction, but only by modulating tendencies that are universal or default" (p. 29). But I would additionally argue that the deepest of these roots of human sociality should not be ascribed to the species at all. Instead, some of the "peculiar emergent properties" (p. 26) of interaction are actually human 'modulations' of entirely independent features of reality. The physics of time and matter is completely independent of life and human existence, yet it pervades everything, allowing us (and indeed forcing us) to experience life as a set of overlapping sequences of events and processes. Basic elements of social interaction such as 'sequence' and 'alternation', though certainly modulated in uniquely human ways, are integral parts of a reality which has existed without us for billions of years.

Thus, while Levinson (2006, p. 46) notes that languages and cultural systems across the world seem to have "eerily similar subsystems" for turn taking, question-answer sequences, and so on, some of these subsystems may emerge this way because they cannot be any other way: events that are contingent on other events entail that a *sequence* will occur, whether we are talking about grains of sand in an hourglass being blocked by the presence of other grains which have not yet dropped, or human responses to various observed contingencies. And when the favored action of any one of a collection of entities is a contingent response to action by other entities, *alternation* is the natural and inescapable outcome. In other words, when the "moves" during interactions are contingent responses to other moves, a pattern of alternation will emerge whether the system involves a collection of colliding spheres, jostling grains of sand, or thinking individuals with free will. This is just one example of how reality imposes inescapable temporal structure on experience. While many of the details may be modulated by culture or cognition or physiology, some basic principles such as 'sequence' and 'alternation' do not need to be learned, nor do they derive from our biological makeup, except in the sense that we are a collection of corporeally separate beings whose actions and interactions we can observe.

I hold a number of additional assumptions about the human experience which influence my approach: I assume that people are in nearly constant thought, with one idea often leading to another in orderly as well as meandering sequences, and with a great deal in the way of incomplete or prematurely cut off fragments of reasoning. And as social beings, we broadcast or conceal our thoughts and intentions and purposes amongst each other, often in pursuit of specific goals, all of which we must attempt to "keep in mind" by way of attending to or somehow monitoring each other and our own thoughts and acts. We must remember or be reminded of what is going on and what has been happening at recent and more distant timescales, keeping ourselves oriented in such a way that our next thoughts and actions will continue along a fruitful course that we are trying to follow at any given moment.[1]

How do we manage it? Clearly there are difficulties: the passage of time brings with it a continuous onslaught of fading memories, distracted attention, disrupted or derailed intentions, broken lines of communication, and so on. The threats to coherent existence seem mostly due to its elements constituting mere moments and finite intervals embedded in this inescapably continuous passage of time. It would not be surprising, therefore, if we had developed certain defenses, and learned or manufactured others: some memories are strong rather than fleeting, and knowledge can be built up and maintained through practice and experience, with some of it accessible at any time and much of it anchored or made accessible via external artifacts such as architecture, objects, writing, and stored data of all kinds (see Hutchins 1995, 2006; Goodwin 2000, 2006). An exploration of the development of these abilities in the complex history of our species is well beyond the scope of this work, but there is certainly plausible pressure for their cultural or biological evolution, because they involve adaptations which surmount some of the facts of existence which might otherwise make survival more difficult.

Zeroing in on humans in face-to-face communication, the same threats to coherence exist: words must end so others can begin, attention is distracted at crucial moments, people try and often fail to attend to and recall some of what has been discussed, or is

[1] There may be times when we feel no pressure or desire to order our thoughts at all, but such moments of total cognitive aimlessness are not normally possible when we are engaged in face-to-face interaction.

being discussed, or will likely be discussed next. We can make things easier through the use of diagrams and other useful contextual clues, both pre-existing and created on the fly. Yet even without external tools, we have the advantage of our own flesh-and-bone artifacts: bodies which last from moment to moment, with articulable, changeable configurations which likewise can be counted on to persist in some fashion even as our thoughts bounce around in constant danger of losing their coherence. Thus with the temporal and physical constraints of our reality come matching pragmatic difficulties, yet also a few affordances literally ready at hand for overcoming or at least struggling against them.

Much previous work in the field of gesture studies has focused on the most energetically expressive moments of gestures, where each achieves its most accented or effortful change in configuration, labeled by Kendon (1980) the gesture 'stroke'. The stroke has since been assumed by many to be the canonical carrier of expressivity or influence (if any) of a gesture's contribution to an utterance. In contrast, my specific aim is to explore the functional possibilities of intervals of gesture *stasis,* in which the hand(s) or body remain in roughly the same position over a span of time, yet are also not simply neutral or at rest. Kendon (1980) named such spans gesture 'holds', and they have not received the same level of intensive study as gesture strokes. As the less noticeable 'ground' against which salient 'figural' strokes occur, the expressive potential of holds remains for the most part unexplored. However, I propose that these spans of embodied stasis afford a powerful array of functions relating to the management and expression of ideas and contexts over time, the exploration of which forms the basis of the chapters to come.

A preliminary example

This example serves to introduce a number of important themes. I first present it in minimal form before revealing its full presentation later in the chapter. In the section of the discourse spanned by the three images below, the speaker has started the task of explaining to his partner how they will arrive in a certain town, represented by the model visible below him in the images. He begins with this utterance:

> "Okay, so here's the plan."

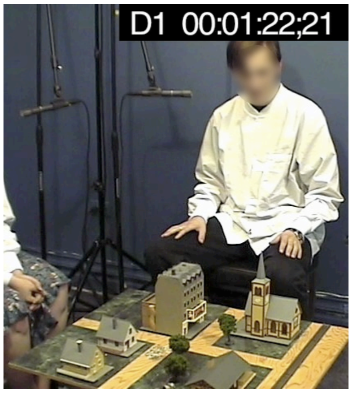
During the two seconds immediately following that introduction, his body takes on the following configurations (eye-gaze is apparent in the unblurred originals but will not be visible here):

1. His head and gaze orient toward what for him is the far left corner of the model.

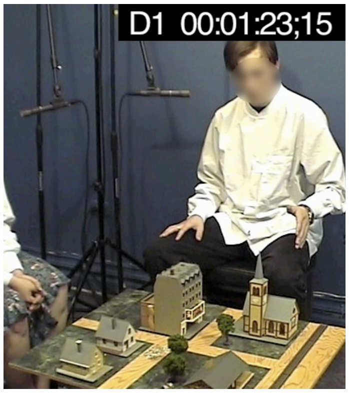
2. The flat, closed-fingered palm of his left hand rises, pointing with its fingertips.

3. Both hands rise, palms facing each other, and shortly thereafter thrust forward.

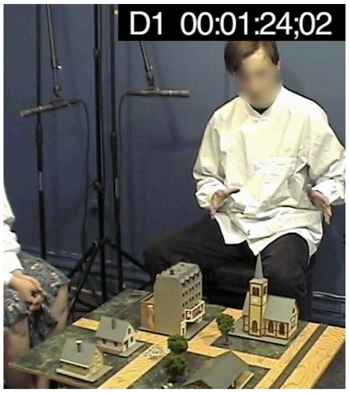
1.



2.



3.



Every example presented here involves *co-speech gesture*, rather than gestures which stand alone. Simply put, co-speech gestures are produced concurrently with speech. But a crucial point of variation involves the details of exactly *when* the different parts of gestures occur with respect to *what* speech. The co-timed speech in this particular example includes the phrase "the train will come in" just after the third image above, coinciding with a forward thrust of his arms. With the addition of this information about his concurrent speech, the situation may at first seem straightforward: his left hand first indicates the path of this train along the tracks as shown in the second image (the tracks are the dark diagonal stripe at the lower right of the frame, visible between two parallel, thicker light-colored stripes). His two hands then thrust forward in time with his speech about the train's arrival.

However, examining the speech in the middle portion of this example reveals tremendous additional complexity. In fact, the phase shown in the second screenshot above occurs during a period of speech disfluency in which he utters "We sh- we'll we'll...," and this interval and its gesture last approximately one second longer—if they would have happened at all—than they likely would have if something like "the train will come in" had been arrived at smoothly instead of as a restart after a hiatus in the flow of fluent speech. At this point, it is useful to interpret his utterance from the point of view of his interlocutor, sitting to his right (i.e., in the left portion of the screenshots, though she is mostly cropped from view here). From her point of view, he first says "Okay, so here's the plan," and turns his head to his left, roughly toward the area of the train tracks. He then begins raising his left hand while stammering "We sh- we'll we'll..." such that his hand is held in position deictically indicating *something* in the general area of the train tracks—which are on a path matching the vector extending from his fingertips—where he and his partner will evidently *do something* (his faltering speech is at least interpretable this far). He maintains this configuration through an additional 0.1s of silence, after which he finally says "The train will come in." During this newly fluent speech his right hand is raised to join his left, a movement which starts on "The," and his hands thrust forward in unison along the vector defined by the tracks (and his previous gesture hold), with the thrust movement beginning just after "will."

During the one second or so of speech breakdown, the listener certainly lacks direct access to the speaker's processes of speech planning and spatio-motoric thinking, but she can still glean rich referential information ahead of—and soon to be integrated with—her perception of the speech and gesture he produces upon regaining fluency.[2] Her attention will likely have been drawn to the relevant portion of the model, following his own shift in head and eye orientation, and the shape of his hand may hint at a path being indicated, perhaps even the train tracks themselves. This information will almost certainly influence the process of her uptake of what comes next—for example, by predisposing her to more efficiently recognize certain already-expected words or grammatical constructions in his speech, or to more readily engage the intended subset of possible semantic structures they can cue, especially with her awareness of the particular portion of the model city which is about to be referred to.

During his disfluency, what might otherwise have been a fleeting moment instead becomes a span of time over which his speech provides a repeated cycle of aborted restarts of variations of "we'll," followed by a moment of silence. Meanwhile, his partner has ample opportunity to analyze what his suspended gesture may be indicating, because the unexpected span provides time over which to analyze the continuously accessible aspects of the scene, including his gesture and body configuration generally. The temporary failure of speech leaves his interlocutor to scrutinize these aspects closely for any hints that might help fill the gap left by the unexpectedly aborted or suspended speech stream. Although we may have an almost analogous 'suspension' in *both* speech and gesture, this event yields a span in which his speech has largely lost its ability to add new information but in which his gesture, on the contrary, has become firmly planted as a persistently accessible artifact and can therefore achieve amplified influence. This is not to say that his disfluent speech loses all function—indeed, in this example, we can surmise that his repetitions of "we'll" may amplify the interlocutor's focus on this element which in turn affects her uptake of the rest of the perceptible activity (e.g., she may begin to wonder "we'll do *what*, exactly?").

Without knowing exactly what she is attending to in this very instance, I still argue that situations like this one afford a great deal of highly referentially significant clues which have clear potential to affect the listener's understanding a great deal, meaning that they very likely *do* affect it in many instances. Even just the fact of a speaker's disfluency has been shown to instantly predispose listeners to assume the speech will be about a certain subset of possible referents. Arnold, Fagnano, and Tanenhaus (2003) and Arnold, Tanenhaus, Altmann, and Fagnano (2004) showed that a production of "thee, uh..." (plus noun) instantly predisposed listeners to look toward unmentioned referents that would constitute "new information," whereas productions of simply "the" (plus noun) predis-

[2] I intend flexible notions here for what I have termed "processes of speech planning and spatio-motoric thinking," but see Slobin (1987, 1996) on the concept of "thinking for speaking," and Kita (2000) regarding "spatio-motoric thinking" in fragmentary speech-gesture utterances like this one.

posed listeners to look toward already mentioned "old information" referents. These preferences were measurable almost instantly, within 200 milliseconds of the onset of the noun (I discuss this work in more detail in Chapter 3). Unless interlocutors somehow cannot make use of visual information at these speeds, it seems unthinkable that they would not be influenced by visual cues, such as information available in gesture, while predicting likely referents. In fact, there is strong evidence that visually observed gestures and actions in general are integrated with linguistic context just as quickly as linguistic cues are (Özyürek et al. 2007; Willems et al. 2007). When disfluency results in a suspension in the progress of an utterance, there is more time to observe gestural cues and also less possibility of relying exclusively on speech. Gesture configurations which span these periods provide a readily perceivable confluence of persistent cues, which maintain a powerful link to the discourse as it existed in the moments leading up to the disfluency, as well to the discourse that is about to resume.

It is worth emphasizing that this mechanism, by which an interlocutor can continue to observe cues relating to an utterance undergoing temporary suspension, *does not* require the gesture to be performed by a speaker in *response* to a disfluency such as a gap in lexical retrieval, to help mitigate the communicative deficit while also facilitating word-finding. We need not agree with Butterworth and Hadar's (1989) belief in such a compensatory impetus, but their suggestion that a gesture available to a listener during a hiatus in speech "will convey at least some of the intended meaning" (p. 173) is quite reasonable if we are careful about our stance regarding the origin of such gestures. During normal speaking, gestures frequently convey complementary information to that which is exactly simultaneous in speech, associating clearly with utterances but not always with individual, isolated words.[3] Gestures performed during the normal course of utterances provide clues to intended meanings that go beyond the content of single words, a fact that becomes all the more stark when the linguistic stream is unexpectedly disrupted and becomes impoverished, while gesture may continue to be observed.

Speech can of course also include features that persist through time, such as pitch, loudness, voice quality or phonation type, rate, and many other things which together form meaningful but often grammar-independent aspects of utterances, informally known by labels such as 'tone of voice' or 'manner of speech', most of which could be subsumed under the term 'prosody'. Gumperz (1982, ch. 5) notes that these are basically variations of the fundamental variables of sound—frequency, amplitude, and duration—overlaid on the simultaneous unfolding of grammatical speech (whose own building blocks are, of course, dependent on these same basic variables). Non-grammatical durational features of the sound stream can tremendously affect the meaning of utterances, for example by evoking and reinforcing the frame or activity or set of assumptions which should take scope over the utterance.[4] Non-verbal durational cues (i.e., produced

[3] I discuss the fallacy of the "lexical affiliate" in greater detail in Chapter 3 (see especially pp. 83-84).

[4] Gumperz (1982) traces these concepts from Bartlett's (1932) *schemas* through Goffman's (1974) and Fillmore's (1976) *frames,* and Levinson's (1978) *activity types.* Gumperz's own *contextualization cue* encompasses the non-grammatical verbal and non-verbal cues which are part of utterances and embed them into such frames, often leading to communication failures since both the cues and the frames are highly culturally dependent and not necessarily shared by people speaking the same grammatical language.

by means other than the vocal tract) can be equally involved in evoking and maintaining many kinds of scope, an ability that will concern us throughout the coming chapters. But the non-verbal cues discussed in my examples will differ from vocal durational cues in at least three respects: (1) With gestures produced alongside spoken language, we do not usually engage the visuo-spatial modality in the task of producing a lexical stream, meaning that co-speech gestures are capable of displaying a rich visuo-spatial configuration across an interval of time while lexical production continues *or* falters. Part of the special capacity of co-speech gesture derives from the fact that it is usually disengaged from the burden of conducting this fine-grained grammatical linearization. (2) The gesture holds discussed here are non-lexical but often display highly specific visuo-spatial information, and thus the potential for highly specific reference, based on mechanisms which are largely absent (or at least much more limited) in the medium of sound, such as spatial deixis (placement and pointing) and iconic imagery. (3) There may also be a fundamental difference in how much can be produced or interpreted simultaneously in the aural versus visuo-spatial modalities, partly because we have multiple spatially separated articulators in the latter (and well developed visual capabilities) but only a *single* point of production, the vocal tract, in the former.[5]

Thus, spoken language is rather dependent on rapid alterations of the sound stream being linearized in time, a process which is easily disrupted. Co-speech gesture, in contrast, while certainly linearized (just as all behavior and experience is necessarily situated in the flow of time) is also frequently suspended or *held* while accompanying both fluent and disfluent speech. A 'frozen' gesture does not suffer the sudden cut-off that occurs when the sound stream is suspended. From the simple fact of enacting a temporary stasis of form across an interval of time, gesture holds can support an enormous variety of natural functions in conversation. In the coming chapters I will demonstrate some of the breadth of these functions, and take a few steps toward categorizing them.

The development of the concepts of gesture 'stroke' and 'hold'

Adam Kendon (1972) undertook one of the first serious studies of the fine-grained temporal structure of gesture movements with respect to speech, and noted that there seemed to be close parallels between (1) the amount of time spanned by a given body configuration, (2) the size of the body parts involved in creating the configuration, and (3) the size of the concurrent unit of spoken discourse. At the broadest level, we have

[5] Signed languages, in which grammar is encoded in the visuo-spatial modality, do in fact make use of multiple simultaneous spatially separated articulators, such as in the many two-handed signs of American Sign Language. Furthermore, as discussed by Liddell (2003), the variable demands for new articulation on the separate components can allow for one element to remain as a "buoy" while additional signing takes place, creating a momentarily lasting cue which may contextualize additional material in a manner very much like what we observe for some gesture holds. Enfield (2009, ch. 5) provides several clear examples of this phenomenon in co-speech gesture, and I discuss it in more detail at the end of Chapter 2.

positionings of the whole arms and body posture, remaining consistent with chunks of spoken discourse of roughly ‘paragraph’ size. With progressively smaller articulators such as the hands and then the fingers, we have progressively shorter configurations timing with smaller units of speech, and naturally the multiple smaller units are nested within the larger ones. Kendon attributes the basic identifiability of these “process units” in body movement and speech, and the *synchrony* between distinctive changes in body configurations and distinctive chunks of speech articulation, to Condon and Ogston (1966, 1967). In a very different paper, Kendon (1980) later defined the idea of a ‘Gesticular Phrase’ consisting of several ‘phases’ including the ‘stroke’, but at the time of the 1972 paper these were not yet codified.[6] The stirrings of the idea of a gesture stroke were at the time quite flexible regarding whether the key feature was a *sharp movement* or the taking up of a *position*, with what would later be called the ‘preparation’ phase plus the stroke referred to simply as “movement to a position” that associates with each phrase of speech (Kendon 1972, p. 200):

> In many instances, it appears to be best to say that the movement that distinguishes each phrase is *movement to a position* that is distinctive for the phrase. Where this is so, we find that this position is reached at the center of the phrase, that is, at the point in the phrase where the most prominent syllable occurs (Hockett, 1958). This position may be held, or there may be a comparatively slow change that follows, or the movement that is to lead to the next distinctive position may be begun. The length of the movement before the position is reached and what happens after this probably depend partly upon the degree of prominence given the phrase as a whole, relative to the other phrases in the cluster to which it belongs.

Here we also see that Kendon was already thinking about “held” positions, but this idea was not yet codified with the term ‘hold’, nor was it set in distinction to the ‘stroke’.

With the later paper, Kendon (1980) introduced his highly influential hierarchy of the temporal structure of gesture. The ‘Gesticular Unit’ or ‘G-Unit’ consists of “the moment [someone] begins the excursion of the limb to the moment when the limb is finally at rest again” (p. 212). A G-Unit contains one or more ‘Gesticular Phrases’ or ‘G-Phrases’, each of which is composed of a set of ‘phases’. These phases include the optional ‘preparation’, consisting of the movement of a limb “from its rest position to a position at which the stroke begins,” the obligatory ‘stroke’ (which by definition every G-Phrase must contain), defined as “a distinct peaking of *effort,*” and the ‘recovery’ or ‘return’ phase, “in which the limb is either moved back to its rest position or in which it

[6] McNeill, Levy, and Pedelty (1990) summarized Kendon’s (1980) definitions while also switching out the word “gesticulation” for “gesture.” In the process, however, they erroneously attributed the 1980 paper’s codified concepts of “G-Unit” and “G-Phrase” to the 1972 paper as well. This mistake was repeated in McNeill’s widely read book (McNeill 1992) and it has been stated again and again throughout the literature ever since. But in fact, the 1972 paper’s more flexible precursor ideas make it well worth revisiting for close examination on its own terms.

is readied for another stroke."[7] The notion of "effort" is here defined in the sense of Laban (Laban and Lawrence 1947; Dell 1970). As clarified later by Kendon (2004, pp. 112, 124), Laban's sense of effort does not necessarily mean a sharp movement. It could also be described as the span in which 'effort' or 'shape' are manifested with greatest "clarity" or the portion of the gesture which is most "well defined" relative to what comes before and after (see also Dell 1970, p. 11, quoted by McNeill 1992, p. 376, which states that exertion of effort may manifest as a change in "tension," "weight," or "focus," as well as movement).

However, at its time of introduction, Kendon (1980, p. 212) crucially stated that "less technically," the stroke could be defined as "a moment of accented movement," and this is the sense that became standard in the minds of most gesture researchers. This, then, would seem to mark the point at which the idea took hold of a stroke *movement* being the obligatory defining feature of most gestures, relegating all non-accented or non-movement phases to optional or auxiliary roles. Without disputing that such phases are optional and less salient and perhaps less obviously expressive than the typical stroke, I would also suggest that many researchers have prematurely disregarded their expressive potential. The notion that the stroke is generally "the most" expressive or contentful part of a gesture has often been rendered instead as a definition describing it as simply "the" expressive or contentful part: "The stroke is the phase that carries the gesture content" (McNeill 1992, p. 84). "The stroke is the meaningful part of the gesture" (Kita 1993, p. 7). "The stroke is the gesture phase with meaning" (McNeill 2005, p. 32). The stroke "constitutes the expressive phase" (Seyfeddinipur 2006, p. 83).

David McNeill (1992) brought Kendon's (1980) hierarchy to a wider audience, and simultaneously introduced a number of changes and additions (largely matching those presented earlier by McNeill, Levy, and Pedelty 1990). 'Preparation' and 'stroke' were brought in unchanged, but the term 'retraction' began to be used in place of Kendon's 'recovery' or 'return', and 'Gesture-unit' and 'Gesture-phrase' now stood for Kendon's 'Gesticular Unit' and 'Gesticular Phrase'. Much more importantly, however, McNeill introduced Kita's (1990) terms 'pre-stroke hold' and 'post-stroke hold' as new optional phases in which the limb(s) or digit(s) engaged in gesturing are found to hold in place before or after the stroke.[8] In McNeill's (1992, 2005) view, the timing of gesture phases

[7] Kendon has more recently (2004, p. 112) adjusted his definition of the 'recovery' phase: it cannot be a phase in which the limb is "readied for another stroke," because that should more properly be labeled a 'preparation' even if it immediately follows a stroke. Instead it is only a phase in which the limb is relaxed and drawn back toward rest position. This matches McNeill's (1992, 2005) definition of the 'retraction' phase. In addition, Kendon (2004) removes the 'recovery' from membership in the G-Phrase of the preceding stroke, although it is still part of the overall G-Unit. The reasoning behind this change is not explained, but it is perhaps meant to capture the idea that when the limb is being withdrawn, the gesture is no longer intended as 'active'.

[8] Before adopting Kita's terms, McNeill et al. (1990) briefly referred to these hold phases before and after the stroke as the "preparation-hold" and "stroke-hold," respectively. As discussed below, the latter term has more recently come to be used for something quite different. It should be noted that no copy of Kita's unpublished 1990 manuscript is currently to be found (Kita and McNeill, personal communication), but its main points are explained quite well by Kita, van Gijn, and van der Hulst (1998), as well as by Duncan (1996), Kita (1993), and McNeill (1992, 2000, 2005).

with speech are especially useful as a "window" into the mind of a speaker, revealing the way imagistic and linguistic information unfold together in a synchronized pulse from a common origin. This common origin is termed a "growth point" (or GP) and is theorized to begin as a 'psychological predicate' (Vygotsky 1986), a basic departure from context that emerges in communication in both the symbolic linguistic stream as well as the imagistic gestural stream, with the two modalities tightly coordinated. The stroke is described as synchronizing with exactly the speech elements that have sprung, with gesture, from the GP, meanwhile the optional pre-stroke and post-stroke holds are claimed to arise precisely because of this synchrony (see the concise discussions in McNeill 2000, and Duncan 1996). First, a pre-stroke hold is claimed to occur when the preparation has been completed, but the stroke must still wait for the speaker to arrive at the appropriate elements in the speech stream with which to synchronize. In this way, the end of the pre-stroke hold is said to reveal exactly the boundary between speech elements which are not emerging from the GP, and those that are. Second, a post-stroke hold is claimed to occur when the stroke is completed before the synchronized speech items have been fully pronounced. Almost opposite to the pre-stroke hold, the post-stroke hold effectively extends the scope of the stroke to encompass the rest of the speech associated with the GP, even though the stroke itself is finished.[9]

McNeill (personal communication) has clarified to me that in his current thinking, the post-stroke hold is therefore definitely meaningful and expressive, much as the stroke is. Meanwhile, he does not believe the pre-stroke hold is intentionally expressive, but would not characterize it as simply meaningless: it is a visible manifestation of the "dawn" of the GP. It "arises when the GP, in existence since the start of preparation, is being unpacked into a construction that has a set locus or loci for the linguistic part(s) of it, and this locus/loci has/have not yet occurred when the preparation movements have come to a halt (most likely because the hands have reached the center of the gesture space: the semiotic zero point)" (David McNeill, personal communication, June 21, 2010). If this process is visible to an analyst of a videotaped conversation, I would argue that it is also conceivably revelatory to interlocutors engaged in ordinary conversation. Some of the imagery of the GP, manifesting as shape or position (if not yet as a stroke movement) can become available to a listener even if the speaker has not yet reached the point of maximum effort or expressiveness. This 'presaging' becomes especially potent if

[9] In earlier writings, the post-stroke hold was even more explicitly identified with the stroke: instead of viewing it as an index to a now finished stroke phase, McNeill (1985, p. 361) referred to the phenomenon as a "static stroke" following a linked "dynamic stroke." Similarly, McNeill (1989, p. 176), while describing something identical to the post-stroke hold, with the identical function of extending "the gesture as an organized expression … through the whole stretch of parallel speech," nevertheless characterized the phenomenon as an extension of the stroke itself. Referring to a gesture accompanying speech consisting of "picks it up," McNeill writes that "even though the stroke movement ceases with 'picks', the gesture hand, without changing shape, may float at the locus reached at the end of the stroke movement, and the stroke therefore last until the particle and object have been uttered." Along similar lines, though with different motivations, Kendon (2004, p. 112) also elected to create a union of stroke and post-stroke hold, for which he coined a new term, the gesture "nucleus," defined as the stroke plus any post-stroke hold (but not pre-stroke hold). The nucleus is "that part of the action that carries the expression or meaning of the gesture phrase." Thus Kendon here clearly identifies post-stroke holds as being carriers (along with strokes) of the semantic content of gestures, but leaves pre-stroke holds out of contention.

the speaker's utterance becomes suspended after the GP's inception, but before the gesture's expression in a stroke (see Chapters 2 and 3, as well as the example discussed in the current chapter).

McNeill's (1992) work was revolutionary in its approach and influence and helped usher in the vibrant modern field of gesture studies as we know it today, especially with regard to its technique of using close video observations to infer the dynamic, unfolding thought processes of speaker-gesturers engaged in producing utterances. However, I believe our horizons could be expanded in at least a couple of areas:

(i) As I have stated, in making statements such as "the stroke is the phase that carries the gesture content" (p. 84), McNeill may have intended that other phases of gesture carry a different kind of content, or carry content dependent on that carried by the stroke, but on the most straightforward reading it implied that non-stroke phases are simply *not* phases that carry gesture content. We should be explicit in avoiding such restrictive uses of words like "meaning," "content," and "expression" and allow for the possibility of expressiveness at moments other than those strictly intended by a speaker. A phase that is not "expressive" from the speaker's point of view may nonetheless reveal referentially significant information to the interlocutor. This is even true of preparation phases, because they can reveal a trajectory toward a future location as well as an inchoate shape (see discussion in Chapter 3). We should similarly avoid the assumption that a pre-stroke hold is barren of semantic content merely because it is waiting for the gesture stroke to synchronize with speech. From an interlocutor's point of view, there is already a readily interpretable locus and perhaps even a shape—the early perception of which could easily affect the process of uptake of the impending gesture stroke and linguistic material. We should not expect speakers to be completely blind to these processes taking place at times other than the *intended* expressive peaks of their utterances.

(ii) The introduction of Kita's (1990) terms refocused the term 'hold' to refer not just to any period of gestural stasis but especially to pre-stroke and post-stroke holds, which were characterized as arising out of the stroke's synchronization requirements. As discussed by Duncan (1996, p. 43), such "syntactic" holds have not been considered to display their own semantic content. Instead, a pre-stroke hold signals that the gesture is "waiting for speech to catch up" (Kita 1990), and a post-stroke hold is basically an index to a previously enacted stroke which bore the appropriate semantic content. These are crucial features of many holds adjacent to strokes (and in particular they help make salient any stroke thus encompassed), but they are just a slice of the wider phenomenon of gestures entering periods of stasis. We should not restrict ourselves to the study of holds from an exclusively speaker-centered perspective within Growth Point Theory.

Along these lines, while making full use of McNeill's (1992) system, Duncan (1996) nonetheless discussed many of its limitations and suggested a number of adjustments and elaborations. First of all, although she several times affirmed the idea of the gesture stroke as the seemingly exclusive meaning-bearing portion of gestures ("The stroke phase is defined as that portion of a gesture where it assumes a semantically interpretable form," p. 27), she also stressed that in her analysis, gesture strokes need not be movement phases. The full expressive extent of a gesture may occur as a type of static presentation rather than "a moment of accented movement" (Kendon 1980, p. 212) because, for example, a durative state of affairs is much more naturally expressed this way. At her time of writing, the canonical view of the stroke as "accented movement" was already well established (in spite of any affordances in Kendon's original technical definition), therefore it is not surprising that Duncan chose to give her stationary stroke the explicit label, 'hold-stroke' (p. 43). This label somehow became reversed to 'stroke hold' in later years, as referenced by McNeill (2005) and McCullough (2005). In any case, Duncan's term reflects both adherence to Kendon's (1980) original strict definition of a Gesticular Phrase, in which a stroke is always present, while also making explicit her finding that the most expressive portion of a gesture does not always involve movement.

Duncan (1996) also discussed Kita's (1990) "syntactic" pre-stroke and post-stroke holds at length, in the process introducing a new kind of post-stroke hold which she termed "semantic" and which, unlike post-stroke holds that merely index a prior stroke, she deemed to carry independent semantic content *as holds:* "a 'semantic' post-stroke hold … is one that bears some sort of evolving semantic relationship to the movement stroke that it follows as well as to the speech with which it, the hold itself, co-occurs" (p. 44). Her example is of a gesture enacting an event of pulling on a suit and keeping the lapels grasped—the hold on the lapels is not simply an index to the pulling-on movement, but rather a direct part of the event referent. This is not simply a movement stroke followed by a separate hold-stroke, because the referent is unified and communicated in a single 'pulse' of speech and gesture. No examples of "semantic" pre-stroke holds were discussed or deemed possible, consistent with the view, in growth point theory, that a pre-stroke hold is held exactly because it is a *suspension* of semantic interaction between speech and gesture. This is also consistent with Kendon's (2004) exclusion of the pre-stroke hold (but not post-stroke hold) from his expressive gesture 'nucleus'.[10]

However, just as Duncan found examples of events naturally composed of a meaningful movement plus meaningful hold, it seems that we should easily find single events composed of a meaningful hold plus meaningful movement. For example, speech about an underwater person coming up for air could be accompanied by a pre-stroke hold showing the hand covering the nose and mouth, followed by a movement phase enacting the release of the hand to allow breathing air at the surface. A momentarily aimed 'gun-hand' gesture followed by a firing motion could be another example. Clearly there can be shape and location information in a pre-stroke hold, because it is the stage immediately before a stroke and these features cannot be expected to form instantaneously at the

[10] Duncan does, however, acknowledge that holds other than 'hold-strokes' (and, presumably, "semantic" post-stroke holds) are themselves "meaningful, but in a different sense and on a different level of linguistic analysis than what is intended" in her work (Duncan 1996, p. 44, footnote 14). She mentions that another kind of hold is the persistent use of one area of gesture space corresponding to discourse organizational units. This kind of location-maintenance is quite similar to the discourse level patterns discussed early on by Kendon (1972), and developed further in more recent writings by Duncan (2008).

stroke: any change in handshape or limb position is subject to inertia, according to the amount of mass being moved.

Perhaps due to similar reasoning, and influenced by Duncan's (1996) examples of the meaning-bearing potential of non-movement phases, Kita, van Gijn, and van der Hulst (1998) proposed a major revision to Kendon's (1980) G-Phrase structure. They take the rare view that *both* pre-stroke and post-stroke holds are officially part of the "expressive" portion of gestures, while also acknowledging that the center of expressivity of a gesture may be a hold rather than a stroke movement. They do not adopt Duncan's (1996) term 'hold-stroke', referring instead to such phases as 'independent holds'. No distinction is made between "syntactic" and "semantic" pre- or post-stroke holds in their model, with all of them grouped as 'dependent holds' adjacent to a stroke, but still part of the 'expressive phase'. The expressive phase, in their model, replaces Kendon's (1980) stroke as the only obligatory part of a 'movement phrase', and it can consist of either an 'independent hold' *or* a stroke with optional pre-stroke and/or post-stroke 'dependent holds' (this means, of course, that a 'movement phrase' does not actually have to involve movement). Their full model is given in Figure 1 below (Kita et al. 1998, p. 27):[11]

Movement Unit = Movement Phrase *(one or more)*
Movement Phrase = (Preparation) ⇒ Expressive Phase ⇒ (Retraction)
Expressive Phase = Independent Hold
Expressive Phase = (Dependent Hold) ⇒ Stroke ⇒ (Dependent Hold)
Preparation = (Liberating Movement) ⇒ Location Preparation >> Hand Internal Preparation
Retraction *(when followed by another movement phrase)* = Partial Retraction

Figure 1. The structure of a phrase of gestural movement, according to Kita et al. (1998)

This model introduces at least three significant revisions to the hierarchy of Kendon (1980) and McNeill (1992). First, it places all 'hold' phases, whether independent or dependent on a stroke, as official members of an expressive phase, thereby avoiding the oft-repeated stance that it is the gesture stroke which is "the" expressive phase. Second, the proposed stages within the preparation phase are highly suggestive of some degree of preliminary semantic interpretability during that phase, both for location ('location preparation') and shape ('hand internal preparation'). Third, Kita et al. note that any hold adjacent to a stroke is structurally ambiguous between a 'dependent hold' linked to that stroke, and an adjacent but 'independent' hold. This hints at the possibility of one hold type transitioning into another without any visible change, which begs the question of whether there is actually a transition at all: could there instead be single holds spanning multiple clauses?

[11] In their rendering, parentheses mean 'optional', the equals sign means 'consists of', the "⇒" symbol indicates a discrete transition, and the ">>" symbol indicates a transition which is "normally blended" but occasionally discrete.

Kita et al. (1998) published their paper in a conference proceedings that may not have been widely circulated, but fortunately an electronic version is now available. McNeill (2005, p. 32) discusses the existence of Duncan's non-movement stroke phases (referred to by McNeill as "stroke holds") as well as Kita et al.'s (1998) identical 'independent holds', but he makes no other mention of the far-reaching revisions Kita et al. attempted in their paper. McNeill also remarks (p. 32, footnote 2) that "stroke holds are not 'holds'," a curious statement that continues the practice of using the term 'hold' only to refer to pre-stroke and post-stroke holds whose existence is governed entirely by the synchronization requirements of the stroke. Seyfeddinipur (2006) also cites Kita et al. (1998) among several other works on the subject of gesture phases, but in the next paragraph states that the stroke "constitutes the expressive phase" (p. 83), a claim that seems rather at odds with Kita et al.'s (1998) model, given that it characterizes the stroke as being just one of several possible elements in an expressive phase.

Now, I would like to stress again that a gestural cue need not be intended as expressive for it to have an effect in conversation. I take a generous view of listeners' capabilities to deduce information on the fly from a variety of cues, and this includes gestures that may not be intentionally expressive on the part of the speaker, but still provide referentially or pragmatically significant information to an interlocutor. My use of the term 'hold' is most akin to Kendon's original (1980) coinage, where it simply meant a period of gestural stasis but did not yet have a strict position within the hierarchy of the gesture phrase.

A radical shift in perspective: McCullough 2005

Most recently, Karl-Erik McCullough (2005) has proposed a far-reaching revision of our concepts of 'stroke' and 'hold', and his approach has strongly influenced my own. His radical shift is to place gesture movement and gesture stasis on equal ground: both are capable of active representation, each of a different character. Gesture strokes, he argues, are expressive manifestations of a change in representation: they embody difference and transformation, the incrementing of information. Gesture holds, meanwhile, express referential continuity and discourse cohesion across stretches of time. Such spans of continuity can be extremely brief, or modulated by the timing of gesture strokes (as in many of the holds that McNeill and Kita have discussed in their models). But there is no reason that referential continuity must be bounded into phrase-sized chunks. For McCullough, any hold is capable of "maintenance of coherent representation over stretches of evolving discourse" (p. 5).

McCullough was motivated to propose this revision based on findings of gesture holds routinely crossing clause boundaries. The approach shares much in common with Smith's (2003) discussion of the utility of "persistent structure" in gestures and manipulable artifacts, any of which may be recruited, depending on available resources, whenever a constant representation is motivated by the discourse. McCullough (2005) also stresses that gesture holds can occur simultaneously with stroke movements. First, as I have mentioned (footnote 5 on p. 7 above), one gesturing limb can hold in place while another performs new gestures. But McCullough also discusses a less obvious but very frequent form of layered strokes and holds, in which small beat movements are superimposed onto held gestures. These miniature stroke-like movements may serve to draw attention to a gesture hold to “realign, or reestablish the immediate relevance of the held representation with the coordinated speech” (McCullough 2005, p. 617). A gesture can be prolonged, while simultaneously used as the gestural focal point of new increments of speech. If we were to shoehorn this phenomenon into the standard version of McNeill’s (1992) Growth Point Theory, we would have to separate each refocusing movement and the surrounding pre-stroke or post-stroke holds into individual gesture phrases. Instead, it is far simpler to allow that an extended hold lasting across multiple clauses can optionally undergo explicit, incremental realignment with speech by way of superimposed beats.

Kendon (2004, pp. 179-180) has also noted the potential utility of maintaining a gesture for an interlocutor across stretches of discourse, but his discussion is limited to a curiously narrow context. He remarks that there are times when a “narrow-gloss,” word-like gesture representing a single concept is performed in conjunction with a word that names that same concept. Because such a gesture does not appear to add any information beyond what is already given in speech, Kendon was motivated to consider that it may perform some other function beyond simply naming the concept it embodies. For example, it could allow a speaker to keep the concept in active presentation during subsequent speech, and to emphasize it and make it more conspicuous, and perhaps prevent a listener from ignoring it. But what is striking about Kendon’s discussion is that this mechanism is presented as a special phenomenon associated with narrow-gloss gestures, when, as McCullough and I are at pains to emphasize, it is common among all forms of gesture representation. There is nothing unique to the mechanics of maintaining a narrow gloss gesture in active representation: *any* manual gesture enjoys the same physical affordances. If anything, gestures that do not represent encapsulated, word-like concepts are actually less restricted in terms of what speech they might take scope over, across multiple clauses of speech.

How to read this dissertation

All of the examples to be discussed here are a product of a specific culture and language community, dealing with a specific task with particular demands, and little variation in physical setting. This usefully narrows the scope of investigation to something almost manageable in a single document, but also clearly brings up the question of which of my claims are plausible only for the specific passage of discourse in question, or for the overall corpus, and which might derive from near-universals of the human species, or even ultimately from physical and temporal facts of reality. Undoubtedly, all such aspects are operative here, and apply in combination in interesting ways. It is useful, therefore, to keep these differences in mind throughout, and it is my hope that some of my suggestions will serve as a jumping-off point for research into very different cultural and linguistic settings, which may reveal some of the same fundamental behaviors and functions while also elaborating on them in completely different ways from those mentioned here. I intend each of my examples as illustrative of functions which may be operative in many different instances, and I cannot claim complete access to the experience of the participants in my recordings. I claim strong *plausibility* for these functions in each token instance presented here, and it is my hope that the reader will in turn become convinced, as I have, that these functions are very likely to operate productively in many real-world interactions.

My examples come from a corpus of dyadic conversations recorded in the McNeill Lab Center for Gesture and Speech Research at the University of Chicago in May-October, 2000. There were 45 separate pairs of participants, and each took part only once. Before each conversation, one participant was isolated from his or her partner for approximately two minutes and allowed to study a small model village while reading a set of instructions. This participant will always be referred to here as **Participant A**, and will always be seen, from the reader's perspective, on the **right side** of the screenshots. Each conversation begins with Participant A explaining the task to **Participant B**, who can be seen on the **left side** of many screenshots. In order to maximize the reader's access to the details of the examples while remaining within the confines of the printed page, I have formatted each example with the following features:

(Please refer to the **Key to Symbols and Annotations**, pp. vi-x)

- A speech transcript which is intended to be highly readable, and free of most annotations unrelated to the task of displaying the relative timing of speech and gesture.

- A linking scheme matching each still image to its corresponding moment in the speech transcript, by highlighting the most closely timed syllable or segment or pause marker, and assigning it a label which is also appended to the image.

- A prose description of the gesture behavior leading to each snapshot image, in order to help the reader recreate the total dynamic scene, including what is taking place *between* the snapshots.

- A basic system of annotation showing the relationship of a few broad gesture phase types to the speech transcript and still images. This includes periods of zero gesture or rest position, periods of active gesture stasis (i.e., 'holds'), clear gesture apexes (i.e., movement 'strokes'), small movement excursions ('beats', see McNeill 1992), movements toward new gestures ('preparation' onsets), and movements toward rest ('retraction' onsets), but omits most other details of the gestures, which can instead be gleaned by looking at the images and reading the prose description.

- Timecode stamped to each image revealing the session number and precise time of the captured frame, as measured from the start of each video.

- Annotations indicating the amount of time (to the closest tenth of a second) elapsed between each sequential image shown. (The same information is available, in less accessible form, by looking at the differences in timecode from image to image.)

Speech rate could not be clearly coded while simultaneously fitting images on the page efficiently, but it can be inferred, very roughly, from the time elapsed between each transcript-linked image. Most importantly, I aim to rectify a major deficiency in most descriptions of the relationship between gesture and speech, which is the almost universal lack of comprehensive sequences of images. Instead of one or two photographs or sketched outlines per example, I follow an approach somewhat similar to Liddell (2003) and include many small sequential photo images for each example, lined up carefully with transcript. The advantage of focusing on holds is that each can be shown with an image near its start and near its end, offering convincing evidence for the span it covers, and showing the amount of change versus stasis across that span.[12]

While examining each example, I urge the reader to attempt to act out the described gestures while speaking or whispering the words. Intonation is left uncoded, but my occasional inclusion of periods, commas, and question marks is intended to indicate rough contours and help make the transcripts more visually accessible. It will of course be impossible for the reader to reproduce exact details of form, but by reenacting a dynamic scene based on the information provided, and considering the experience of the simulated speaker or listener, many of my arguments based on the relative timing of speech and gesture will become clear. Thus it is my hope that the reader will be able to use his or her own body as an informal basis for exploring and assessing my arguments, and follow with more in-depth investigations according to individual area of expertise.

Example 1.1 – the chapter's preliminary example, in full annotated form

See pp. vi-x at the start of the thesis for an explanatory key to the annotations and layout shared by all examples. In the example below, note that each of the indexed shaded items in the full speech transcript, at the top of the example, is replicated in the partial transcript reprinted below the images. The transcript below the images is stretched out in order to line up each shaded item as closely as possible to a vertical line running down the center of each image, and the corresponding alphanumeric index is placed above each image, always on that same vertical line. Because the transcript below the images is stretched out horizontally, less of it fits on the page. Therefore, take note that while all of the intended transcript is given at the top of the example, the version reprinted below the images will frequently have material shaved off from the beginning and end, and occasionally from the middle as well. These alterations to the transcript are necessary in order to present as many images as possible, at a size that takes advantage of every millimeter of horizontal space available. Because of this maximization of the use of images, and the fact that sequential images span variable temporal intervals, it was not possible to use a constant temporal scale for the horizontal dimension. The dynamically changing rate of speech must be inferred from the time values given between each image.

[12] The lack of permanently embedded, easily activated video clips in this and nearly all other written works is naturally an even bigger deficiency than a deficit of images. Permanently embedded videos will no doubt become standard in coming years as we do more and more of our reading electronically, but for the time being I have opted not to attempt official inclusion of supplemental video files.

The presentation has certain affinities with Kendon's (1972, p. 201) "movement flow-chart," such as the fact that in the gesture annotation, a solid horizontal line indicates a body part being held still (either as a gesture hold or at rest). But I have explicitly lined up the annotation with readable transcript, rather than abstract bars indicating presence of speech. Kendon's flowchart offers no easy way to know the details of the gesture movements, and my image sequences similarly are limited in their ability to describe movement (because they are merely snapshots of instances in the process). However, the images work very well to show the temporal span of static configurations, since I have consistently presented the snapshots so as to show both the earliest and last moment of a held configuration. The movements themselves can largely be extrapolated well enough, for my purposes, by simply interpolating between the snapshots.

I have also purposefully left out detailed information on loudness and intonation contours such as we see in Kendon's (1980, p. 214) second wonderfully detailed diagram. Including such marks, without any way of recreating the true recorded contour for the reader, could cause an exaggerated or otherwise inaccurate rendition that is no more useful than the rendition achieved without any such markings. For a classic discussion of the various biases inherent in any choice of transcription methods, see Ochs (1979).

a b1 b2 b3 c

(1.1) Okay · so here's the plan. · # · We sh* we'll* we'll* ·· the · train will come in

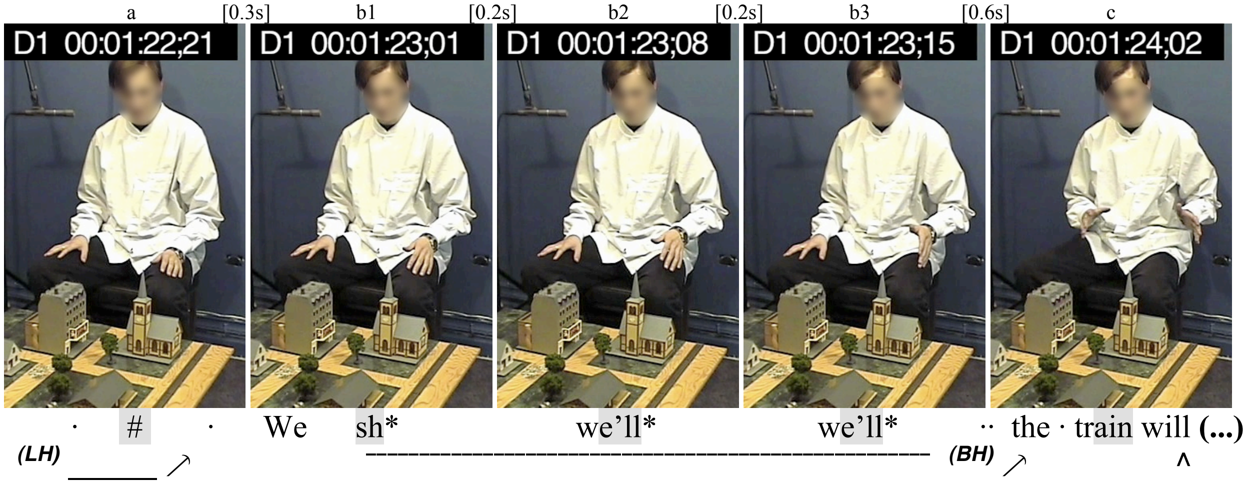
a: *BH begin at rest on thighs.*

a-b1: *LH rises slightly from thigh and holds in place, subtly beginning to indicate the path of the train tracks (visible as a dark diagonal stripe at the lower-right of each image).*

b1-b3: *LH stays almost in the same position through the disfluency, but gradually rises an additional inch as the palm slowly rotates to vertical, solidifying into a more clearly defined deictic even as he fails to restart fluent speech.*

b3-c: *Coinciding with the start of 0.1 sec. of silence, at* [23;18] *LH breaks from its hold and begins moving up and to his left by a few inches, nearly simultaneous with RH rising to join it, which occurs beginning at* [23;21] *timed exactly with the onset of "the" (which marks a return to fluent speech, disrupted only by a slight hesitation before the word "train").*

c: *BH have now finished the preparation phase for a BH forward-thrusting stroke which will time with the end of the word "will." By the end of "come in," BH have returned to rest (not shown).*


a
[0.3s]
b1
[0.2s]
b2
[0.2s]
b3
[0.6s]
c
D1 00:01:22;21
D1 00:01:23;01
D1 00:01:23;08
D1 00:01:23;15
D1 00:01:24;02
· # · We sh* we'll* we'll* ·· the · train will (...)
(LH) ______ ↗ -- (BH) ↗ ^

The images in this example fall into three natural groupings: first, we see the participant's physical state while at rest, with both palms on his thighs (image [a]). Next, three images show his almost stationary, but very slowly rising and rotating LH palm, suspended in mid-air during his disfluent speech (images [b1] to [b3]). Finally, one image shows his state just before he thrusts BH forward in a stroke (image [c]). Additional images could be included, such as his state just after the stroke, or his state after returning to rest, but in this example my primary focus is on the period leading up to his return to fluent speech, and the inclusion of additional images would require a reduction in the clarity of presentation for the rest. In addition, in this case the stroke is sudden but travels a short distance, making it difficult to show in static images, and furthermore his state upon returning to rest is not significantly different from his state in image [a]. The difficulty of portraying punctual movements such as gesture strokes is an unavoidable weakness of the printed page (and readers with access to the original video clips are encouraged to refer to them often). Held gestures, however, are more easily represented in still images. A recurring template I employ will be to present at least the following four snapshots relating to any given held gesture: an image just before the hold (as with image [a] here), an image at the very start of the hold (image [b1] here), an image at the very end of the hold (image [b3] here), and an image just after the end of the hold (image [c] here). This allows for quick assessment of the temporal span of the hold (whose duration can be gleaned from the time values given above the images), as well as a clear presentation of what it overlaps with in the speech stream.

In this particular case the configuration of the held gesture evolves noticeably, but the rate at which it evolves is much slower than the comparatively punctual movements just before and after, hence I have chosen to code it here as a 'hold' using a superscore line in the gesture annotation tier below the images. It is not always possible to decide whether something is a 'truly' held gesture, nor is it necessary to do so. Rather, I wish to highlight aspects of the utterance which persist across time and multiple speech units, and allow that certain elements may not remain completely static (such as the precise angle of the participant's wrist in this example). In fact, it would be quite unnatural for a gesturer's total performance to remain completely static across any length of time. The many independently articulable parts of the body instead exhibit a complex layering of dynamically changing configurations that overlap with each other in time, some of which are very short or relatively 'punctual', others of which are broad and span larger intervals.[13]

Returning to the example at hand, the gesture annotation of "(LH)" indicates that the markings at the bottom of the example refer to the participant's left hand only, until, reading to the right, the marking "(BH)" is reached, at which point the rest of the markings refer to both hands in unison. Depending on the example, separate lines of gesture annotation may be stacked to represent separate articulators, when relevant, and these may be between participants (e.g. they could be labeled "A's RH" versus "B's LH"). In

[13] What I intend by this is that there are distinctive contrasts of *relatively* punctual versus durative configurations. Obviously any action takes place across a measurable interval and nothing is truly a singularity in time. The point is that behaviors occur on a variety of overlapping timescales, including those which can span across multiple instances of shorter, comparatively 'punctual' events. Such events include the apex of gesture beats and most strokes, and individual syllables of speech.

other examples there may be no indicator of which hands are being marked, such as when the images and prose explanation are self-explanatory and the handedness does not change during the example.

The rising arrow below the first image in this example is placed just to the right of the shaded inhalation marker, indicating that image [a] captures the very last moment at which his hand is at rest before being lifted. Meanwhile, the superscore line beginning at the shaded segment below image [b1] indicates that the image shows the first moment at which his hand arrests in place and becomes a 'hold'. The blank section in between is the preparation phase, whose onset is marked by the arrow, and whose duration, in this case, coincides with the 0.3 seconds given above the vertical white line between images [a] and [b1]. No downward arrow retraction phase is given, because the held LH gesture moves directly into the BH gesture shown in image [c]. It is important to note that the relative positions of the images and the items in the gesture versus speech tiers are what is crucial, rather than absolute horizontal distance on the page. As I have mentioned, due to space constraints, varying rates of speech and orthographic word size, and the varying amount of time elapsed between each chosen snapshot, I have not opted to arrange the images, transcript, and gesture annotation on a timeline with precise horizontal scale. Instead, time values are given above the images and in timecode, with the part of the speech transcript corresponding to each image placed as close to its center line as possible. The variable visual density of intervening material simply follows suit. Thus in this example, a large amount of space is taken up by images [b1]-[b3] and their associated transcript and underlying 'hold' annotation, but image [b3] occurs only 14 frames after image [b1]. In contrast, image [c] occurs 17 frames after image [b3]. This variation in horizontal scale is not intended to mislead—it is simply an artifact of the arrangement of equal-sized snapshots occurring at unequal intervals, and the time annotations are intended to help clarify the actual temporal relationships.

In the prose description above the images, each section is often headed by a range of alphanumeric indexes rather than a single index. This is simply because one of the goals of this section is to reconstruct a dynamic scene shown only in a series of snapshots, and therefore must refer to the spans between the images rather than only to the images themselves. Occasionally, precise timecode values will be given in this description, in order to pin down specific events occurring in between the timecode values shown in the images. In general these timecode values will be given in seconds and frames only, omitting minutes and hours.

Making sense of the corpus

Many of the points I wish to make may be convincingly evident in the examples themselves, even without detailed knowledge of the experimental context or the idiosyncratic sources of the conversational topics being negotiated. Nevertheless, the reader will obviously gain a fuller understanding with the aid of some background information.

Common factors underlying the data

The videotaped conversations discussed here were recorded during the months of May through October, 2000, as part of the NSF/KDI Project of the University of Chicago and Wright State University / Virginia Tech, under the supervision of David McNeill and Nobuhiro Furuyama at the McNeill Lab (Center for Gesture and Speech Research) at the University of Chicago. One of the goals of the original project was to develop a tool for investigating the real-time three-dimensional position of the head and limbs of pairs of participants as they held a conversation. This involved the use of five simultaneous MiniDV digital camcorders, with a dedicated pair for each participant (and a fifth which recorded a zoomed-in view of one of their faces). The images in the present study draw primarily from just one of these recorded angles, occasionally making use of another when appropriate.

As an undergraduate research assistant, I recruited the participants, operated the cameras, and administered the procedure. Participants responded to an advertisement that was posted as a flyer, and also as a message sent to the "subject pool" email list, a resource administered by the Psychology Department at the University of Chicago. The advertisement stated that participants were invited to sign up as pairs for a study on "communication," in which they would be videotaped and compensated with $10 each for 30 minutes of their time. Most of the 45 pairs of participants were undergraduates at the University of Chicago; several were grad students and staff. Upon arrival at the lab and following consent procedures, participants were asked to choose who would be the "primary subject." While his or her partner waited in an adjacent room, this "primary" participant (hereafter designated "Participant A") read a sheet of instructions while studying a model village. In the first 32 sessions, Participant A did this preparatory work while seated in the same spot he or she would occupy during the actual recording, which took place with the model village present. An overhead schematic of the model village, including this seating arrangement, is given in Figure 2 below. The sheet of instructions is given in Figure 3.

In sessions 33-45, the model was kept in an adjacent room from where the recording took place, and Participant A examined the instructions while viewing the model from the same position as in the other sessions (see Figure 2). Upon returning to the recording room, the participants then sat in the same arrangement as in the other sessions, but without the model village. In these later sessions, Participant B never glimpsed the model and had to rely entirely on Participant A's description.

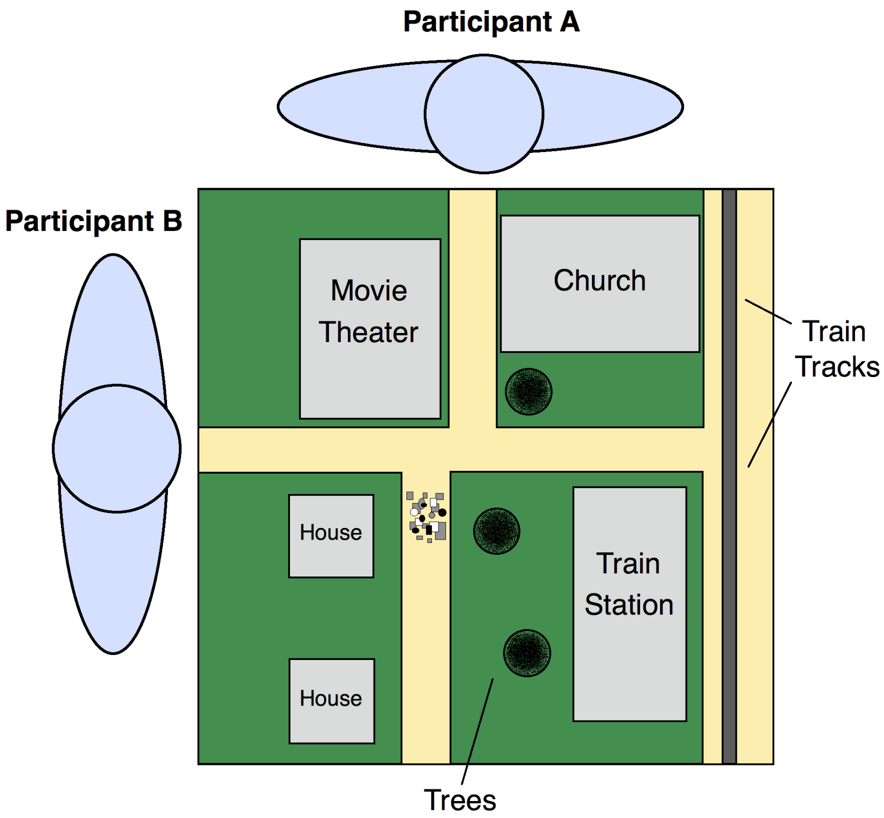


Figure 2. Overhead schematic of the plastic model village, with study participants

> A family of intelligent wombats has taken up residence in an abandoned movie theater in the town of Arlee. You and your assistant need to catch the wombats so that you can send them back to Australia. You will be taking the train to Arlee, so be ready to get off at the station right after you pass a church on your right. When you get off the train, go around the station and cut though the adjacent park to meet your assistants: pass between the two trees and you should reach the house, number 33. Then go next door and ask the neighbors in 35 (you'll notice the road construction in front of the house) to assist you. The movie theater is across the intersecting street. One of you should go in the front entrance and scare the wombats out the back entrance. With the help of the people in 33 and 35, you should be able to snare the wombats as they exit the rear entrance. Explain the task to your assistant and decide on who does what, and what equipment you will need to bring with you.

Figure 3. Sheet of instructions, shown only to Participant A

As can be seen in the many images captured from the video recordings, each of the buildings of the village was a relatively detailed plastic model with many features not shown in Figure 2 above, and the participants did not view the model from directly overhead but rather from their seats. However, the diagram should at least help clarify the layout that was readily apparent to Participant A as she or he read over the instructions before each recording session. Upon Participant A's signal to the investigator of his or her readiness to begin the task, Participant B was seated, and the participants were encouraged to "go into detail" and discuss freely for about 10 minutes. After switching on the cameras, a movie-style clapper was used to officially initiate the conversation, and the investigator monitored the proceedings from a separate room until the participants decided the conversation was over. Most conversations lasted 10-15 minutes; the longest was 25 minutes.

Evaluating the scope of the data, and what we can draw from it

I want to stress that the specifics of the participants' instructed topic are not crucial to my arguments. The admittedly unusual scenario allowed for some consistency across the 45 dyads recorded, but only in terms of the overarching topic that participants knew they were supposed to attempt to discuss. This instructed topic was one facet of a larger set of situational pressures stemming from the cameras, the model city, the odd labcoats they were made to wear without explanation (it was in fact an externally imposed engineering requirement for better isolation of their hands in the video stream), and of course the institutional setting of a paid Psychology Department experiment. At the University of Chicago, signing up for such experiments serves as a frequent site for social excursions by groups of inquisitive undergraduates. The upshot is that nearly all participants took the task "seriously" at least most of the time, but also remained relaxed and surprisingly unselfconscious, with plenty of digressions, joking, friendly sparring, and so on. The instructions allowed participants a great deal of freedom in how to approach the task, and they exhibited tremendous individual variation in their level of interest or boredom, cooperativeness (with each other) versus antagonism or competition, and commitment to realist versus absurdist choices in their detailed plans of action. Knowing only that we wished to study "communication," their speech and gesture naturally also displayed a tremendous amount of individual variation.

While this particular stimulus scenario may have favored certain phenomena over others that might have occurred in different contexts, the corpus nevertheless exhibits tremendous breadth. The limited scope of the data already provides far more variety than can be adequately addressed in the scope of a single study. Without consciously intending it, I find that my approach matches that of Kendon regarding the examples in his (2004) book, and his statement applies just as well to my examples: "We have made no attempt to stratify or systematize the sampling of participants. Examples have been selected for the clarity with which they illustrate different kinds of gesture usage. Questions about how gesture usage might vary systematically by age, sex, setting, discourse circumstance and the like, although of great potential interest and importance, have not been explored" (p. 110). My corpus of recorded "Wombats" sessions is numbered D1 through D45 (these identifiers are included in the timecode stamped to each image), and while I did survey the entire corpus, the twenty-five examples presented in these chapters come from just nine of the sessions: seven from D1, six from D3, one from D8, one from D13, four from D20, one from D34, one from D36, three from D38, and one from D45. An index of the examples, including the timecode encompassed (in minutes and seconds) and a brief quote of some distinguishing transcript, is given at the start of the thesis to aid in easy reference (p. v).

In no way do I presume that the particular spread and apparent prevalences among the observations given here will closely mirror a statistically sound survey of human gestural behavior generally; but I *will* argue that each of the phenomena categorized here is likely to occur (whether rarely or frequently) in a wider variety of conversational settings than the one in which my recordings were made, and across languages and cultures as well, of course with myriad local modifications. These local modifications are of great interest, but we should also try to tease out, throughout any investigation, those facets which likely derive from more universal truths about the human experience. The typology described here is a subset of the undoubtedly richer spectrum of human gesturality in general, but it is nonetheless a subset that I expect to be generalizable to some degree, and observable in a wide variety of manifestations beyond this limited corpus.

Investigations like mine, which present and discuss single examples of documented behavior, rather than more targeted statistical measurements under highly controlled conditions, serve us best when they inspire scholars with completely different training to conduct their own research using their own approaches. While my examples come from a laboratory setting, my approach is more akin to Enfield (2009) and Kendon (2004) in that I treat the recordings as normal conversations which happen to take place in idiosyncratic circumstances. It can of course be said that *all* conversations take place in their own unique circumstances, and the fact of being physically situated in a laboratory should not brand my corpus as completely ‘unnatural’ by mere association with much more restrictive laboratory settings. The participants are well at ease with each other and the cameras, and in any case cannot help but utilize the normal, embodied communicative abilities which they bring to any context of interaction. That said, what they choose to talk about is of course heavily directed by the setup and the instructions and their beliefs about the institutional expectations. Depending on the presence of the model village, their gesture and speech are also to varying degrees “environmentally coupled” to it (Hutchins 1995, 2006; see also Goodwin 2000, 2006, 2007). But these, in fact, are the ‘local modifications’ characteristic of any instance of interaction that occurs in the world—that is to say, quite natural.

My project is to convince the reader that certain functions for gesture holds are very likely, which I attempt to accomplish by *showing* them in action such that the reader can simulate the scene and experience a version of it at a distance, using some of the same abilities as those required to read these paragraphs and communicate in daily life. In this instance, having the language of study be the same as the language of exposition is a great help, as is the fact that the setting of the recorded conversations is easily explained and easily understood. My claims are certainly not meant as the final word—this would be absurd given the very limited scope of both my methods and the recording context—but rather an invitation to others to conduct their own research on the behaviors in question. This includes the much more targeted controlled experiments favored by psychologists and psycholinguists, as well as the much more ethnographically comprehensive investigations within other contexts and cultures ‘in the wild’, as favored by linguistic anthropologists.

David McNeill acknowledges (personal communication) that his own work has tended to focus on how a close, frame-by-frame analysis of video recordings can reveal the dynamic thought processes of a speaker-gesturer, rather than the dynamic interpretive experience of an addressee in situ. He views a passage of speech and gesture evidence as ultimately revealing the most about the person doing the speaking and gesturing, because the behavior is unequivocally connected to that person and the behavior has unequivocally occurred. The thought processes of a speaker's interlocutors, in contrast, are a great unknown, because there is usually no way of knowing how they are interpreting the speaker's behavior, or what they are attending to or failing to attend, or any details about the dynamic course of their evolving comprehension. My approach here is almost exactly complementary: rather than search only for the external manifestations of the millisecond-timed thought processes of a speaker, via his or her synchronized body movements caught on tape, I treat myself as both analyst and as a kind of listener or hidden interlocutor. My approach includes the fundamental assumption given eloquently by Kendon (2004) in the first sentence of his book: "Willingly or not, humans, when in co-presence, continuously inform one another about their intentions, interests, feelings and ideas by means of visible bodily action." While speaker-focused evidence yields incontrovertible real-time facts about the speaker, a defining characteristic of being in co-presence as communicating humans is that interactants, like analysts, are constantly attempting to guess the speaker's evolving mental processes. This simulation of 'the other' is one of the fundamental building blocks of the human "interaction engine" (Levinson 2006).

The speaker's thought processes are partly unknowable from the evidence at hand, and so are those of all the interlocutors in the recordings. Yet by using my native competence as a somewhat removed listener, with the ability to play back various portions repeatedly and observe them at natural speeds, I can surmise *plausible* interpretations and processes of referential influence for a person engaged in communication with the speaker. I am not trying to prove what the listeners on the tape *actually* comprehended, but rather what some of their *likely* interpretations and thought processes may have been, or what the experience could be for others observing a similar behavior in a range of contexts. I know a given interpretation is plausible when I am convinced of experiencing it myself; the remaining task is to argue the case to you, the reader, and attempt to illustrate the features of the example in such a way that you can experience some of it yourself as well. While usually not stated this way explicitly, this is in fact the approach used by any linguist or gesture researcher who presents examples and expects them to be interpretable, and it pervades such major recent works as that of Kendon (2004) and Enfield (2009).

The notion of a 'functional typology'

Underpinning my approach is a principle which I have again found, coincidentally, highlighted in Kendon's work (2004, p. 225). This is the idea of a *typology of functions* operating on a single gesture form. Kendon points out that a gesture can interact with variations in speech and other details of context such that many different functions can be served by the same gesture form, depending on the details of each instance. These can include "referential" functions as well as "pragmatic" functions (with the latter subdivided by Kendon into "performative," "modal," and "parsing" functions). Kendon illustrates this by showing that a particular gesture form, in this case an open palm directed downward or obliquely, can be understood as an act of rejection (a performative function), as a negative intensifier to an evaluative statement (a modal function), or to mark the end of the course of a line of argument (a discourse parsing function). Such functions can manifest individually or in combination. Unifying them into a "gesture family" associated with the single gesture form is a what Kendon calls a *"semantic theme,"* a theme which in this case is "the interruption or cutting off of some process or line of action that is in progress" (p. 226).

Along these lines, this entire dissertation focuses on an even more minimal specification of gesture form—the phenomenon of gesture stasis across spans of time. Among fully realized gestures satisfying this specification, the examples show that very similar-looking gesture holds appear to support a variety of functions, both referential and pragmatic. Referential functions, as Kendon (2004, pp. 159-160) discusses, manifest in gesture in two overarching ways: *representation*, and *pointing*. Representation can occur via 'modeling', in which a body part serves as the physical model of something, 'enactment' (and pantomime) in which the action of the gesture reproduces some aspects of an action or event being referred to, and 'depiction', in which the gesture movements or configuration outline the shape of something, whether the boundaries of an object or the curve of a path. It is important to emphasize that these avenues of representation are not mutually exclusive, and they are also not independent of pointing functions: any gesture occurs at a locus and draws attention to itself even if it does not point to a distant location and also draw attention away from itself, as occurs with many pointing gestures. Similarly, no pointing gesture is free of representation, as the shape and manner of execution are instrumental in recognizing a pointing gesture's intended vector for directing attention. Another way of describing all this is to say that gesture's referential capacity relies on the two core principles of *iconicity* (representation as described here) and *deixis* (pointing or directing of attention), and as Duncan (1996, p. 23) emphasizes, every iconic gesture is in some sense also deictic, just as every deictic gesture is in some sense also iconic. A good example is a path depiction, which requires the depictive representation of the shape of a path, but also frequently does so at a distance, unequivocally involving a deictic pointing function to trace the moving position of an entity following the path. Any depiction, in fact, can be said to be composed of moving deictically pinpointed positions, since deixis and pointing need not occur at a distance.

But is there a "semantic theme" that groups together the functions of gesture holds, which simply share the property of holding in place across intervals of time? As McCullough (2005) notes, holds fundamentally involve *maintenance* of a configuration across time. This "theme" may seem somewhat vacuous—after all, it is basically the definition of a gesture hold. But it actually highlights a number of problems in existing gesture typologies, such as the requirement, even in Kita et al.'s (1998) update on the definition of pre-stroke, post-stroke, and "independent" holds, that there must always be a dividing line between the gestures of one speech-gesture composite and another. To place an invisible boundary in the middle of a continuously held configuration seems to add more events to the production process (and certainly the observing process) than are warranted. By emphasizing the semantic theme of "maintenance," or we can refocus on searching for functions such as representational maintenance, discourse level maintenance, speech act maintenance, and so on, instead of interpreting holds exclusively as stroke-bounding markers of short-term production processes such as the Growth Point. I present evidence for a variety of such functions in the chapters below.

'Retrospective' and 'prospective' cues in composite utterances

Throughout the examples, of continual importance is our understanding of the human experience of time, especially the flow and sequencing of thoughts and actions. Applied to conversational situations, instances of gesture 'holds' synchronizing with, or failing to synchronize with, the spans of more or less co-timed speech and joint conversational projects (see Clark 1996), we are faced with a particular level of temporal granularity that we might informally refer to as 'conversational time'. This is a timescale at which we can discuss the relative onset and offset of the different elements that combine to form composite conversational "moves" (Enfield 2009; Kendon 2004; and cf. "composite signals," Clark 1996), as well as the temporal relation of these moves to each other and to other features of context. Enfield (2009) coins the term *enchrony* to designate this temporal realm, as distinguished from the much broader notion of *diachrony* (a term typically referring to developments taking place across historical time).

Communicative acts do not occur in a vacuum: the participants will have observed and taken part in relevant shared and unshared experiences leading up to the moment in question; i.e., there is context, including 'what just happened'. The acts themselves are not temporal singularities, but instead take place across intervals of variable duration often loosely conceived of by the experiencers as 'now'; thus there is also the question of 'what is happening' and how it relates to (and participates in building on) the ongoing features of context. And of course, communicative actions occur before, and are goal-oriented toward, an onrushing future including 'what may happen next'.

Writing in the 1930s, Schutz (1967, 1973) characterized any action as having a "because motive," which is the completed (and thus, for Schutz, factually "objective") stimulus or state of affairs which has led up to an action, as well as an "in-order-to motive," which is the subjective purpose the actor has in mind, reaching toward an intended outcome (Schutz 1973, pp. 67-72). The because motive is not merely an in-order-to motive as it enters the past: it is the solidified state of affairs a prior in-order-to motive has helped shape, and which is now being responded to. In keeping this distinction Schutz emphasizes the human experience of time as *not* being simply a progression along a mathematical timeline, with future and past treated equally. Indeed the state of affairs achieved may frequently differ starkly from the state of affairs intended.

Another way to characterize "because motive" and "in-order-to motive" is to talk about *response* and *initiative* (Linell 1998), as applied to conversational moves. Growing out of the tradition of turn-based Conversation Analysis (Schegloff 1968, Sacks et al. 1974), every step in an interaction can be judged in terms of its responsive aspects (i.e., the way it grows out of the because motive), as well as its initiatory aspects (those which strive for an intended next state of affairs). These are not necessarily separable: in fact very few, if any, moves are responsive without being initiatory, just as very few, if any, moves can be thought of as initiatory without also being responsive to some state of affairs. More than anything, the terminology simply emphasizes that any human interaction, and its internal structure of "moves," is situated in time and involves a sequence—and at any given moment in this sequence, the actor is responsive to the conditions and stimuli leading up to and encompassing that moment, while also initiatory toward intended futures.

Enfield (2009, p. 9), meanwhile, characterizes an action's immediate past as a "stimulus," "cause," or set of "conditions," and calls the following step the "response" or the "effect." Note the difference here between Enfield's terminology and that of Schutz: the in-order-to motive is never a "response" or an "effect," although it always involves the intent to bring one about. This is merely a matter of perspective, of course: in calling the future item a "response/effect," Enfield is characterizing it from a further-future perspective—that is, one in which the in-order-to motive has led to something which has occurred. Enfield presents the model in a three-stage diagram with (A) stimulus/cause, (B) the speech and/or gesture in question, and (C) the response/effect, but I would argue that the terms for A and C fit better into a traditional two-stage (adjacency pair) analysis, because of the retrospective point of view they all assume. To illustrate: although Enfield's middle stage (B) occurs before what he labels the "response/effect" (C), it is certainly already the response or effect of the stimulus/cause/conditions occupying the first stage (A). Similarly, although his middle stage (B) occurs after the stage labeled the "stimulus/cause" (A), it is this middle stage, not the first stage, which is the immediate stimulus/cause/conditions of the response/effect occupying the last stage in the diagram (C). Thus, Enfield's three-stage diagram is really an amalgam of two identical two-stage diagrams, with the middle stage being both a response/effect as well as a stimulus/cause. This multiplicity of functions is, of course, the quintessential truth of the simultaneous *response-initiative* function of every conversational move (cf. Linell 1998), thus there is nothing inaccurate about Enfield's diagram, though he somewhat misleadingly describes the "response/effect" stage in parallel with Schutz's "in-order-to motive."

The more detailed discussion by Hanks (1983, p. 54) helps to demystify this confusion: we of course take both kinds of perspectives on future events and actions, the forward-looking "in-order-to" motives which (at least in part) drive our actions, as well as the expectation that there will be outcomes, which the parties involved can then reflect on and evaluate after the fact. While intentions do not translate directly into outcomes, there are enough socially mediated means of achieving conventionally intended ends that Hanks groups these patterns as a third essential element that mediates between intentions and outcomes. These "socially conventionalized ends" form the core of our understanding of how to link intentions with outcomes. Speech Act Theory (Austin 1962) is essentially an attempt at describing our socially mediated categorization of these strategies, and our means of evaluating the various ways they succeed or fail. With this more complete picture, in analyzing an utterance or any other action, the temporal context we must take into account includes the prior conditions ("because motive"), the ends intended ("in-order-to motive") or actually realized (when analyzing after the fact) or expected based on conventionalized types, and whether some of these elements are treated as being within the scope of the 'now' of the utterance—that is, subject to "simultaneous interdependence" with various other co-occurring and ongoing features of context (Hanks 1983, p. 64). These notions are as powerfully realized in gesture as in any other realm of activity. Their true force, however, becomes clear when gestures are considered in their normal role as part of *composite* utterances combining speech and other kinds of signs, because with composites come the possibility of varying temporal alignments of the component parts.

As I illustrate in the coming chapters, gestures are particularly adept at persisting beyond, and/or beginning before, some of the evidently co-expressive speech elements with which they combine during utterances. One way to characterize this could be to claim that such gestures are actually concrete, physically manifested cues directly demonstrating the "because motive" or "in-order-to motive"—not in an abstract philosophical sense, but as undeniable, perceivable stimuli which should be taken into account in any model of utterance production and comprehension. To summarize the main points of this line of thinking:

Because Motive — prior conditions and established ground

The visuo-spatial medium is particularly suited (compared to, say, the medium of sound) to the creation of artifacts lasting across spans of time, and our bodies, via gesture performance and other means, provide the most direct way of engaging this medium. Gesture 'holds' instantly create lasting (if still temporary) material anchors. Like other "persistent structures" (Smith 2003), they are solidified artifacts of prior conditions, maintained into the present and available to renewed attentional focus. Simply through continued salient presence, the artifact of a previously enacted gesture (and thus, occasionally, part of a previous utterance or conversational "move") can literally *perform*, in the present, part of the "because motive" of the current utterance. To illustrate diagrammatically, at Time 1 in Figure 4 below, during a stream of speech morphemes, a gesture is first enacted. While additional morphemes of speech are articulated (whether as part of the same utterance or continuing into subsequent utterances), the gesture configuration is held and maintained, such that at Time 2, it persists as a still-salient 'retrospective' reminder of prior conditions, namely those at Time 1, or those of the entire interval between Time 1 and Time 2.

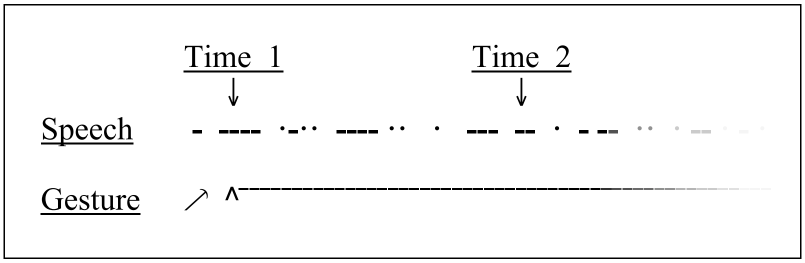


Figure 4. A 'retrospective' gestural artifact (at Time 2) of the speech and gesture at Time 1

In-order-to Motive — clues to an intended future

Switching perspectives now, from one in which a gesture has been maintained for a span of time, to one where the gesture is just now being enacted, consider a case where the speaker-gesturer performs speech and gesture which, although relating to the same referent(s), are nonetheless produced significantly apart from each other in time.[14] In those cases in which gesture and coreferential speech are not exactly synchronized, the gesture has a tendency to be performed before the speech, and not the reverse (see Chapter 2 below; Kendon 2004). I will discuss some of the reasons for this elsewhere, but for now simply consider Figure 5 below, in which a gesture is initiated at Time 1 in the absence of speech. The gesture, held through to Time 2, is then joined by coreferential speech, thus completing (in this simplified example) all the elements of the composite utterance. But at Time 2, although the gesture is arguably part of the "now" of the utterance (since it is coreferential with speech taking place at that time), the fact of its being available to perception and interpretation ever since Time 1 means that the process of interpretation, as well as production, may be quite different than it would have been without the early gesture. The participants will have been *primed* by this "presaging" gesture. If we consider the moment of full interpretability of the utterance to be the (later) time at which speech and gesture are brought together, then at Time 1 the speaker-gesturer can be said to be physically performing part of their "in-order-to motive": whether purposefully or not, certain details about the impending utterance may be revealed, 'prospectively', through gesture.

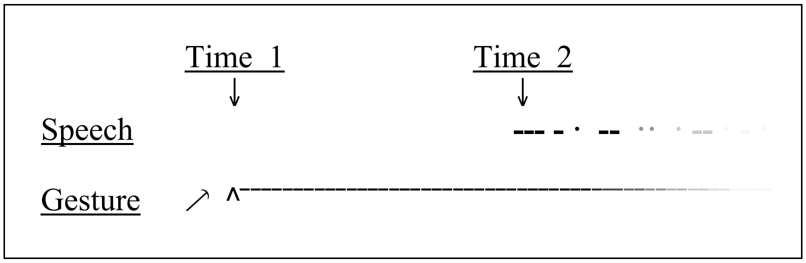


Figure 5. A 'prospective' gestural cue (at Time 1) to the speech and gesture at Time 2

Conventional means — gesture heuristics and expected outcomes

In spite of much emphasis on its novel, uncodified nature in comparison to speech, I have nonetheless mentioned (p. 26) at least two conventionalized, perhaps even cognitively innate mechanisms which are always operative: spatial deixis (pointing), a strategy of

[14] By "significantly apart," I mean any time difference large enough for the earlier event to possibly affect the later event, in terms of utterance formulation or comprehension. Roughly speaking, any time difference greater than about 0.2 seconds (200 milliseconds) will be past the threshold of the reaction time of college-age individuals for both visual and auditory stimuli (see Welford 1980), which provides one possible measure for two stimuli being "significantly apart."

indicating and directing attentional orientation in space; and iconicity, a strategy of representing imagery and physical structure through the metonymic use of the body's articulators as schematic building blocks and outliners. Therefore, although co-speech gesture lacks a grammatical code on the order of spoken and signed language, tokens of gesture behavior do not completely resist categorization. People engaged in conversation observe gestures (of others as well as their own) during conversation and automatically, if unconsciously, assess them along at least these lines: 'does this gesture direct or draw my attention?' (deixis) and 'does this gesture's shape/position/movement schematically represent some physical structure or action I am familiar with?' (iconicity). Relating to Hanks's (1983) characterization, principles like these are then the essential 'glue' by which gestures, like other kinds of signs, help us conventionally link intended ends with accomplished outcomes, applicable for the internal thought processes of the speaker-gesturer as well as intersubjectively, in the evaluation of the communicative efficacy of composite utterances.

Simultaneous interdependence — constructing the static out of the dynamic

Gestures and body configurations engaged across spans of time also provide a direct means of physically performing the scope of the 'now'. Using the body as a metrical device, areas of space as well as spans of time can be *bracketed* in the service of establishing and maintaining an intended topical scope across multiple clause boundaries (see McCullough 2005 for detailed discussion).[15] Although such bracketings must last across a span of time, and therefore have a necessarily non-static timecourse, the temporary constancy of the bracketing gesture may be an active means of 'conceptualizing as static', or of actively resisting the temporality of the interval and instead treating it as a unified object, which could help us in the task of parsing out the moment to moment flow of interaction into discrete chunks such as 'topics'. This is especially true when the beginnings and ends of such spans are also demarcated by a variety of salient boundaries, such as whole shifts in posture, the head, or the upper arms (see discussion in Kendon 1972), spoken discourse markers, and the salient transformation of any held gesture into something else. Static behaviors, including held gestures and postures, can thus provide a means of 'doing topicality'.[16] From a Schutzian perspective, this could be described as a performance which attempts to collapse its own temporal context: the prior conditions ("because motive") of an ongoing bracketing hold are also the intended future conditions (which cue the "in-order-to motive"). The active unification of near-past with near-future, and the bodily performance of this metric, is essential to any attempt at taking control of the variable scope of the 'now'. In Figure 6 below, at Time 1 the speaker-gesturer is not yet actively using the body to bracket a scope over the speech. The interval spanned by Time 2–Time 3 is bracketed by a constant held gesture or posture, during which the near-past and (intended) near-future are actively treated as equal, gesturally. At Time 4, the

[15] I thank Michael Silverstein for suggesting that this can be usefully described as a kind of metricalization.

[16] See Sacks (1989) for variations on this sense of 'doing', e.g., "doing questions," "doing dialectic," "doing intimacy," "doing trouble," and so on.

speaker is no longer actively performing this scope. The end of a bracketing gesture does not automatically end the topicality, of course—it just means that the speaker is no longer 'doing topicality' by gestural means.

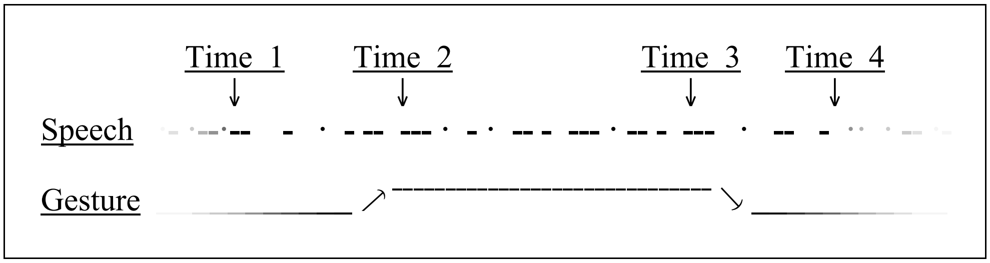


Figure 6. Gestural bracketing: identity of 'prospective' and 'retrospective' during Time 2 – Time 3

Discussion

The reader will note that the diagrams presented in this section have the capacity to morph into one another—for example, the diagram under "because motive," intended to represent the retrospective capacity of a gesture hold, can just as easily serve as the first half of the diagram above representing the bracketing function. Indeed, these perspectives are not at all mutually exclusive. Even a gesture instantiated first with a punctual stroke, timed closely with a highly specific lexeme, can live on as a post-stroke hold and take on topic-bracketing functions. Because of the human cognitive capacity for metonymy, a highly specific referent linked to a gesture does not preclude the same gesture form from also representing the topical scope within which that referent is situated. As McCullough (2005) illustrates so thoroughly, post-stroke holds are simply not constrained to terminate with the end of the linguistic material of a given clause-level "growth point." McNeill (personal communication) points out that such extended holds often appear to include new muscular effort that shifts their position slightly just as new pulses of communication emerge, meaning that they are actually new gestures. Even models such as Kita, van Gijn, and van der Hulst's (1998) would similarly attempt to apply a boundary between a post-stroke hold and any subsequent 'independent hold'. But these perspectives then ignore the continuity of the gesture form, both as a resource for cross-clausal coherence as well as its source. This is what McCullough (2005) has attempted to capture by referring to the "self-generating" capacity of physically embodied lasting cues such as gesture holds, a capacity I would argue derives ultimately from the configurational longevity of solid matter in general, when compared with other vehicles of semiosis.

Returning to the theme of Schutz's "in-order-to motive," the reader will note above that I characterized some gesture behaviors as providing "clues to an intended future." I now wish to emphasize that, contrary to the explicit assumptions in Schutz's terminology, we should not be so quick to assume that people are always engaged in a purposeful project with an intended future. This may be a philosophical debate which I have no hope of resolving here, but I find it plausible that we may never know with certainty, from ordinary audio/video recordings, which parts of any given action are truly "projective" in the sense of being directly in support of a purposeful project, versus which are emergent, reflexive, unconsciously habitual, or otherwise not directly planned. However, removing the certainty of 'purpose' from the equation does not do fatal damage, because emergent behaviors still have the power to influence later outcomes: "projective" or not, consciously enacted or not, gestural cues still have the power to affect an addressee's orientation toward impending utterances, as well as affect a speaker-gesturer's own utterance formulations.[17] We should remain open to the idea of consciously *or* unconsciously enacted cues causing conscious *or* unconscious predictions or priming.

In order to embrace this purpose-neutral stance, I will use terms such as 'prospective' and 'presaging', rather than 'projective', 'preemptive', 'anticipatory', and so on. The term 'prospective' is of course meant as a counterpart to 'retrospective', a choice which works well for the kind of evidence I will discuss: I am primarily concerned with exploring what are, essentially, *perceptible cues* to what just was, or what may soon be. Retrospective aspects of a perceptible cue are those which provide a link back to something earlier, whereas prospective aspects are those which, though the cue itself has been enacted, offer hints to something that may be about to occur. In other words, since in either case the cue is able to be reacted to or reflected on, '-spection' seems appropriate as the root of both terms. A retrospective gestural cue of course does not present the totality of the "retrospective reasons" (Hanks 1983, p. 54) for a communicative act, nor does a prospective cue reveal the full set of "projective goals"; rather, such cues are *available* to perception by observers and performers and will have varying degrees of influence.

Conclusion

I began the chapter with a discussion of the inescapable physical and temporal constraints governing human experience, and suggested that one behavior known to occur during face-to-face interactions, the phenomenon of embodied stasis known as the gesture 'hold', may be well suited for functional adaptation. I also discussed the history of the

[17] Some authors take a very narrow view of the possible functions of gesture and believe it is primarily involved in the process of lexical retrieval by a speaker (Krauss et al. 2000, inter alia), while others have primarily focused on communicative possibilities (Kendon 2004 and earlier work). However, there is no reason why these positions must be diametrically opposed. All action occurs in a context, but also generates it: lexical retrieval is obviously also affected by preceding speech, yet we surely do not claim that words occur primarily for the purpose of lexical retrieval and are not communicative. Similarly, highly communicative gestural cues should also be expected to affect ongoing processes on many levels, including utterance formulation and ongoing conceptualizations. Kita (2000) argues for the utility of gesture as a spatio-motoric tool for negotiating difficult event descriptions, and Goldin-Meadow (2003, 2006, inter alia) argues for the capacity of gesture to concretize 'liminal' conceptualizations into full ones. McCullough (2005) rejects Krauss et al.'s insistence on a low-level, lexeme-internal function for gesture affecting lexical retrieval, but readily acknowledges that like all ongoing semiosis, gestures (and perhaps especially held gesture configurations) that serve topic- or referent-maintenance functions will necessarily influence low-level lexical retrieval processes too.

concepts of ‘stroke’ and ‘hold’, and asserted that traditional accounts have not adequately explored the variety of functions afforded by the latter. As I will illustrate in much greater detail, gesture holds are capable of revealing or maintaining access to referentially significant information at both earlier and later times than when it is revealed in speech, even when the two modalities are closely coordinated. In this chapter I also provided details on the structure and annotation scheme that each of my examples will utilize, and I discussed the origins of the video corpus from which the examples are drawn, including common factors underlying this data, and an assessment of its breadth and limitations. I then outlined my motivation for referring to gestural cues as ‘retrospective’ and/or ‘prospective’, when considering a fine-grained comparison of their timing relative to concurrently produced speech.

Outline of the chapters to come

In each of the following chapters I illustrate a particular perspective on the themes presented thus far. Chapter 2 presents a survey of gesture holds coinciding with periods of speech disfluency, illustrating that gestures are capable of continuing the ‘delivery’ of an utterance in spite of a breakdown in speech. I illustrate differences and similarities between the communicative utility of ‘retrospective’ and ‘prospective’ gesture holds, and also discuss the phenomenon of gestural ‘buoys’ lasting through extended sequences of fluent and disfluent utterances. In Chapter 3, I compare my findings to other work on the subject of gesture and disfluency, and bring in psycholinguistic evidence in support of my claims of listener uptake of gestural cues during disfluency or hiatus, such as what occurs when speech is suspended following preliminary utterance commitments. I also discuss the issue of complementarity of expression between co-timed speech and gesture, and how it affects the perennial debate regarding speech-gesture synchrony versus asynchrony. Chapter 4 is dedicated to the phenomenon of gesture holds ‘bridging’ across interruptions caused by outside interference, such as competitive and collaborative speech by interlocutors. I also explore the potential effects on attention and memory, for both speakers and interlocutors, of gestural cues lasting across spans of time during discourse. In Chapter 5, I discuss gesture holds coinciding with, and closely coordinating with, turn transitions between speakers, for which gesture holds appear to be employed to maintain retrospective reference that also enforces a context for speech by another. In Chapter 6, I conclude with a brief synthesis of the material presented in the preceding chapters, and suggest directions for future research.

2 When a hiatus is not a hiatus: Holds during pauses and disfluencies in speech

When we converse with one another, our utterances invariably contain many irregularities that make them different from what might be called 'ideal' or completely fluent productions. To claim an utterance has a disruption in fluency necessarily involves the claim that during some interval of time, speech is instead non-fluent or missing entirely. Regardless of our criteria for what counts as "disfluent," we typically find that such intervals are surrounded by fluent speech on both sides, meaning they represent a *hiatus* in fluent speech. Assuming we can pinpoint the boundaries, a hiatus in speech can therefore be diagramed as three distinct phases (see Levelt 1983, 1989), which Clark (1996, p. 258) refers to as the "disruption schema." I reproduce his diagram in Figure 7.

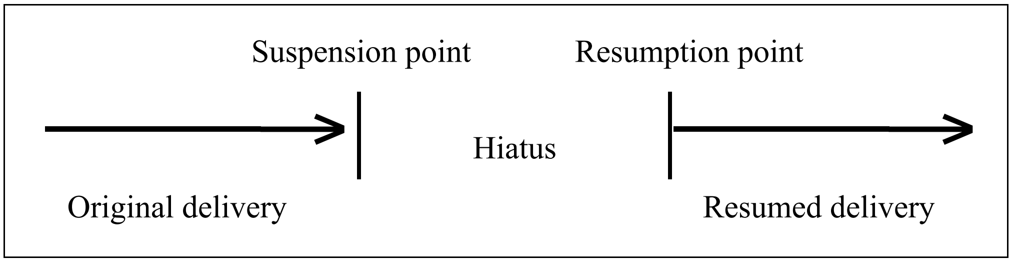


Figure 7. A basic "disruption schema" for utterances, according to Clark (1996)

A speech hiatus can be several seconds long, or almost immeasurably short. It can consist of a total absence of vocal tract sounds, but also frequently gets filled by "uh" or "um," or intakes of breath, or swallowing, or tongue-clicks, or short phrases like "you know" and "I mean," or an extension of a preceding vowel or consonant from before the hiatus. When speakers begin restarting and aborting speech repeatedly, it can be difficult to say with precision what speech should count as a true "resumed delivery" versus repetitive filler of an ongoing multi-stage hiatus (and for our purposes such ambiguities can be left unresolved). It is often assumed that a hiatus in speech equates to a hiatus in an utterance's 'delivery' of content, at least in terms of the content associated with the fluently delivered part of the utterance. However, in this chapter I will show that even disfluent gesture behavior can continue to support 'delivery', as it were, straight through intervals of speech hiatus. This remarkable ability is made possible by three main sets of facts:

(i) **Physical affordances and constraints of each modality.** Any visually perceivable configuration, whether static or mobile, continues to reflect light and support the perception of multiple elements. A complex configuration of the sort we find in conversation is often one consisting of gesture, other current artifacts recruited for use, and the various elements of the environment. All of this yields a large number of potentially relevant elements to be focused on even if gesture articulation freezes or is otherwise disrupted. In contrast, a disrupted sound stream, particularly one originating from the single source of a human vocal tract, cannot continue yielding information unless its frequencies are allowed to modulate in time.[1] Since spoken linguistic material depends on linearized modulation of sound frequencies, verbally delivered linguistic content is thoroughly disrupted during a hiatus in speech.

(ii) **Retrospective imagery maintaining links to prior moments.** Anything causing the speech stream to freeze into stasis or cut-off can *also* apply to the rest of the body, causing gesture to become arrested in place temporarily. Somewhat separately, there are also frequent pauses that occur between speech constituents and are not disfluent in the 'speech failure' sense, and in this type of hiatus we often find extended gesture holds of the post-stroke variety. Regardless, a hiatus consisting of vocal silence or disruption can coincide with maintenance of the immediately prior gesture configuration, which is something more than gestural 'silence' because it embodies a salient *retrospective* link to the gesture and speech of earlier moments.

(iii) **Prospective imagery available at the start of utterances.** Embedded in context, gestural content is often referentially significant,[2] sometimes of meaning that is not directly encoded grammatically. Besides its embedding within global and sequential context, gesture is also deployed simultaneously with speech that presents relevantly related information but in "linear-segmented" grammatical form, rendering the gestural content transparently interpretable. However, because of their differing modes of delivery, the basic imagery is often in place in gesture before the linear segmentation of the related spoken material is complete.[3] Thus, the intermodal timing is irreducibly complex: even though gesture and speech co-contextualize each

[1] An unchanging pattern of sound may certainly be functionally distinct from silence, and a prolonged speech sound originating from the immediately preceding lexeme *might* extend that lexeme's perceivable salience in the utterance, but I would argue that any effect of this sort is severely limited compared to the potential richness of visuospatial content available in a static configuration or scene.

[2] By "referentially significant," I mean gestures whose iconic form/performance or deictic vector are capable of aiding in reference resolution in context (see Kendon 2004, chs. 9 and 10).

[3] McNeill (1992, p. 248) believes this is because the gesture constitutes an idiosyncratic, holistic differentiation of a thought from existing context, while the linguistic material "is a linear-segmented version of the image plus those obligatory linguistic elements required by the standards of good form in the language."

> other and the gesture stroke is generally synchronized to a key stressed syllable in the linguistic stream, the whole linear segmented portion usually takes more *time.* Unavoidably, some of its elements are pre-contextualized by the holistic imagery in gesture even when the two streams are thoroughly synchronized. While this effect is scarcely noticeable in many utterances, it can become quite stark when there is a hiatus in the delivery of speech. For one thing, gesture is not constrained to such strict linearization requirements as speech, and the form of a gesture is often well-defined prior to the stroke. This means that during spans of disfluency, gestures can still yield some of the basic content they were in the process of providing at the time of the hiatus, co-expressive with the now absent speech and serving as a clue to the missing content. Furthermore, speakers do not always withhold an incipient gesture's stroke in order to avoid performing it without speech; some gestures achieve an effortful peak even as the spoken part of the utterance falters. The upshot is that during disfluencies, gestures can reveal strong *prospective* clues about the intended utterance—prospective because, until the arrival of a resumption or repair of the speech to go with the already realized gesture, the utterance is not yet complete. For the duration of the hiatus, such gestures are thrust into the role of supporting a much higher proportion of the interpretable content of the utterance.

Item (i) above concerns the basic properties of any visual, vocal, and auditory system coupled with the physical properties of light and sound. Items (ii) and (iii) concern the relative, intermodal timing of gesture and speech content when a speech hiatus occurs. When a hiatus occurs after a gesture and its synchronized speech have already occurred, a remaining gesture hold is *retrospective* in character, but if the hiatus occurs while a gesture is being performed and involves a disruption of the intended concurrent speech, the gesture is *prospective* in character, and it often holds in place until the speech can resume or be repaired. Retrospective and prospective aspects are not mutually exclusive, because speech and gesture are embedded in a sequential context and a gesture hold can simultaneously maintain, through temporal contiguity, an indexical link to a prior moment while also setting the stage for the next.[4] Examples of all these types are presented below.

Example 2.1 — embodied stasis as a mechanical response

First of all, I wish to show something akin to the pure 'mechanical' body response that appears to be responsible for many gesture holds coinciding with speech disfluency. The gestures in this example do not seem to present any referential content related to the utterance, though they may help intensify certain parts of the message by metaphorically 'picking it up' for presentation to the listener (this is a case of the ubiquitous "conduit

[4] See Sidnell (2005, pp. 79-81) for an example involving similar timing issues: a gesture performed at the start of a disfluent utterance is kept in place and elaborated on with additional speech. It is both retrospective of the intent of the abortive utterance, and also forward-contextualizing during the utterance's reformulation.

metaphor"—see Reddy 1979). First, the speaker performs three brief gestures (not shown), each consisting of a an up-and-down excursion with no hold, quite similar to that shown in images [a] through [c]. These are timed with the stressed syllables of the words "study," "menacing," and "take* bring," with the stroke of the third appearing to hesitate very slightly on the aborted word before reaching its full exertion on "bring." This first hesitation is too subtle to show well via snapshots, but a much clearer case occurs moments later. The first gesture phrase shown in the snapshots (images [a] through [c]) coincides with fluent speech and differs from the preceding cases only in that it is 'waggled'. The next gesture phrase begins just before the speaker's hands have returned fully to rest, and this time a speech hiatus occurs after the word "intelligent," lasting one second until the first syllable of "fauna." During this hiatus, the waggling energy ceases, the hands hold in place while sagging slightly, and then begin waggling and rising again just ahead of the resumption of fluent speech. Thus, coinciding with the hiatus in speech, the speaker's body also appears to enter a sort of stasis, in which the limbs pause in mid-air until he resumes 'normal delivery'.

(2.1) **B**: Why do we need to catch the wombats? ······ *(1.5 sec. pause)*

A: To study them! ···

And · 'cause they're menacing the town, y'know. · You don't wanna let* #

We have to take* bring them back to Australia. · Y'know. #

a b c d1 d2 d3 e

Where* with the rest of the intelligent · # aah fauna · residing in Australia. #

____________↗ ^ ↘ __↗ ‾‾‾‾‾‾‾‾‾‾ ^↘_____

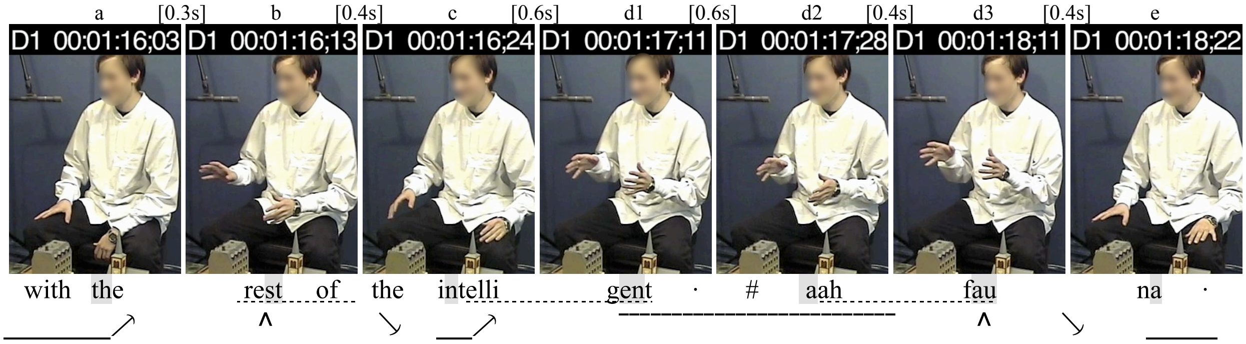
a-c: *BH rise for a waggling stroke and immediately fall nearly all the way to rest again (LH reaches rests back on thigh, while RH is an inch or two above the thigh at its lowest point).*

c-d1: *BH then rise again and waggle in place, with a loose arc of effort rather than a clear stroke.*

d1-d2: *Coinciding with the pause and breath, BH stop waggling and sag slightly while holding in place. At this same instant, his gaze drops from his interlocutor's face (this is obscured by blurring).*

d2-d3: *BH resume waggling and begin rising during "aah," with a weak stroke at the start of "fauna."*

d3-e: *BH return fully to rest by the middle of the second syllable of "fauna," which is also the exact instant at which his gaze returns to his interlocutor's face.*



Unlike with most of the other examples in this chapter, this hold does not necessarily help maintain any referential expressivity through the hiatus, but it illustrates several other important features that occur frequently in disfluency-spanning holds. The maintained gesture coincides with an incomplete utterance that the speaker intends to continue. He averts his gaze from his interlocutor's face exactly during this hiatus, then looks at her again as he drops his hands and finishes his utterance fluently. This is significant because gaze directed at a listener's eyes is one of several factors in a speaker's behavior which can mobilize a response (Stivers and Rossano 2010, Kendon 1967). His partner may interpret the averted gaze, with hands left in place, as a sign of his intention to continue without interference. We do not need to claim that he intentionally leaves his hands in place to broadcast such a signal—rather, it is worth investigating whether this is a sort of embodied stasis resulting from his thought process experiencing a certain kind of interruption. The visible manifestation may stem from basic, inescapable principles: it is fundamentally a suspension in time of a dynamic system's evolving process. Yet this fundamental nature, coupled with ease of detection by other parties, would also make the phenomenon universally available for functional adaptation. It will have become interwoven as part of the evidence we use to guess at each other's thought processes.

The hold in Example 2.1 is also consistent with traditional claims regarding the timing of holds as developed by McNeill and Kita in Growth Point Theory (McNeill 1992, 2000, 2005). There are two alternatives here: if we think of the initial surge at image [d1] as a stroke, then the hiatus is spanned by a post-stroke hold which does not release until the relevant linguistic portion succeeds in being formulated and enunciated (in this case, the noun of the noun phrase). If instead we think of the stroke on "fauna" as the only stroke, then the gesture beforehand is a preparation and pre-stroke hold, which must wait to synchronize its stroke with the proper piece of linguistic material. In either case, Growth Point Theory would state that since the disfluency in speech lengthens the timespan of the apparently co-emergent linguistic constituent, the connected gesture phrase is also lengthened. Alternatively, the phrase could consist of two Growth Points, one which is never completed and therefore includes a gesture literally left hanging (i.e., "the intelligent…"), and the other which is formed when the speaker decides how to complete the utterance and then generates a new exertion of gesture, taking the held position as its starting point (roughly on "aah fauna").

However, other examples presented in this chapter may be more problematic for Growth Point Theory's standard claims about the role of gesture holds, since they include strokes during silence leading to holds during speech, and (as in many of McCullough's [2005] examples) post-stroke holds that are left in place across clause boundaries rather than being dropped when the Growth Point is "discharged." While I remain agnostic about Growth Point Theory as a production model, I will focus instead on the possible *communicative* effects of the gestures in my examples. Beyond the possible floor-holding function illustrated by Example 2.1, the examples in the rest of the chapter illustrate gestures that, from the point of view of the listener, may indeed support continued 'delivery' of referentially significant cues during a speech hiatus.

Holds supporting prospective reference across disfluencies

Recall Example 1.1 from Chapter 1 (pp. 17-20), in which I argued that although the gesture coinciding with the disfluent span may have merely been an aborted preparation, it nonetheless had the power to draw attention deictically to a particular area of the model village, presaging the scope of reference that the newly fluent utterance then assumed. Iconically presented information can similarly hint at reference across the span of a speech failure and ahead of the fully fluent part of an utterance, and occasionally it even suffices in cases where the spoken part of the utterance never actually regains fluency. Such is the case in the next example.

Example 2.2 — maintaining gesture and averting gaze while abandoning speech

In the immediate discourse context leading up to the disfluent utterance in this example, Participant A has just asserted that "wombats aren't that big." His next utterance is an unambiguous continuation of that theme, such that his gesture confirms his intended meaning with enough clarity that his interlocutor responds to the utterance as if it were a completely normal one. The gesture's basic form would arguably have been just as appropriate alongside fluent speech, but in this instance it proves completely sufficient alongside the extended disfluency, and his partner's response suggests she had no difficulty understanding his full meaning. The size-range depicted by his gesture, coupled with "they're" in speech, creates a completely interpretable meaning even though he never succeeds in producing more than this minimal speech.

The eye-gaze of the participants is once again significant in the example, though the reader must take my word for it given that their faces are blurred. Participant B keeps her head and gaze pointed consistently at her partner's face throughout his long disfluency, then averts them away from him at exactly the moment of her response. Participant A, meanwhile, looks at his partner during his previous utterance ("Wombats aren't that big"), to which he expects a response, but averts his head and eyes during his disfluency, especially during the long silent hiatus in the middle. Examples 2.1 and 2.2 are quite consistent with Kendon's (1967, p. 60) early observations on the role of eye-gaze in the maintenance of turns:

> It is suggested that the speaker, by looking at the auditor, signals to him that he is ready for him to start speaking, as well as being able to see whether this signal has been received. In looking away, the other person signals that he has accepted the 'offer' of a change of role. During long utterances it is also found that the speaker looks at the auditor during passages of fluent speech and at the end of phrases but that he looks away during passages of unfluent speech or during hesitations. In this way the speaker can request attention signals from the auditor and, in looking away, can gain time for planning what he has to say, by forestalling any attempt to speak by the auditor.

(2.2) **A**: If I see any coming I’ll just close the door. ··

B: Still there’s always the possibility that they might break the doors down, or #

A: Wombats aren’t that big. ···

B: I don’t know I’ve never met a wombat. #

a They’re* they’re* · they’re* they’re* c1 ·· c2 ·· ·· ·· c3 they* ·· c4 they* d ·· |**B**: e Okay.

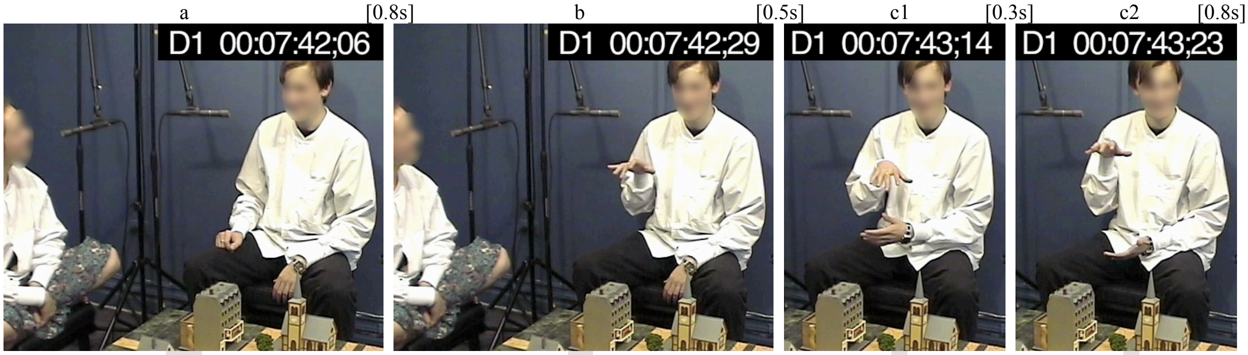
A: They’re* they’re* · they’**re*** they’re* ·· ·· ·· ·· they* ·· they* ·· |**B**: Okay.

a: *Participant A begins disfluent speech and averts gaze toward his left; Participant B gazes at him.*

a-b: *RH begins rising from thigh at* [42;19] *and waggles in place, palm-down with spread fingers, through continued disfluency, with emphasizing upward strokes timing with the word “they’re.”*

b-c4: *LH begins rising at* [43;06], *just as speech is temporarily abandoned, and begins waggling up and down, palm-up in opposition to RH, creating a vertical scale that varies between 1-2 feet. During this interlude he turns his head to his left (furthest at* [c2]*) and back again while his gaze flicks even further in that direction, and then back forward again with his head.*

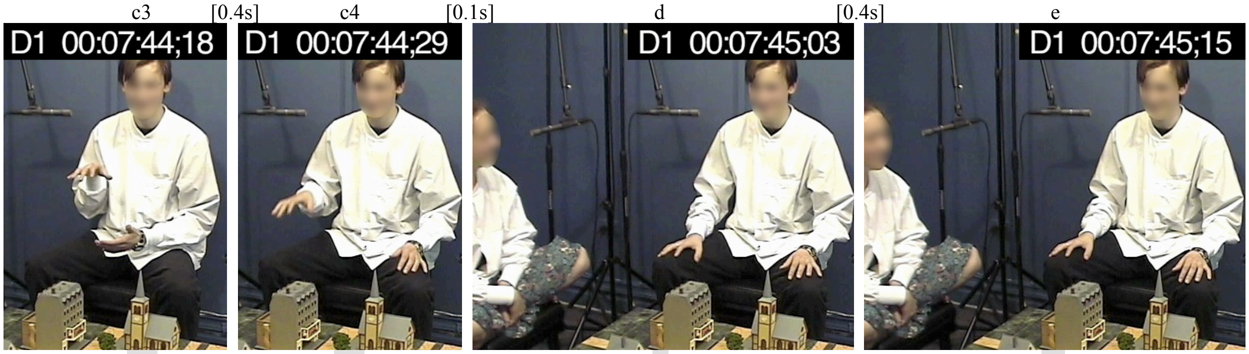
c4-e: *With the last of two restarts, BH drop decisively to thighs. Within 0.5 sec. his partner responds while averting gaze and nodding her head (she continues nodding through 3 sec. of silence).*



A: They’re* they’re* · they’re* they’re* ·· ·· ··

(RH) ________________ ↗ ^ ---------- ^ ------------------------ ^ ------

(LH) ______________________________ ↗ ------------------------



A: ·· they* ·· they* ·· |**B**: Okay.

(RH) -- ^ ------------ ↘ ______________________________

(LH) ---------------- ↘ ______________________________

The presence of listener-directed gaze does not automatically entail a request for the other to start speaking, if Kendon is right that such gaze also occurs at the end of phrases in the middle of utterances. This was evident in Example 2.1 above, where it occurred at the word "fauna" and before the final clause, "residing in Australia." Part of this difference is certainly that the linguistic material should also be consistent, lexico-syntactically, with the end of a turn (De Ruiter, Mitterer, and Enfield 2006). Meanwhile, in Example 2.2, Participant A's action of dropping his hands to his thighs appears to relinquish his literal hold on the floor, such that it is appropriate for Participant B to respond at that moment even though Participant A has not directed his gaze at her. Interestingly, his lexico-syntax also does not indicate the end of a turn (though it doesn't indicate a continuation either), and he satisfies none of the four response-mobilizing criteria named by Stivers and Rossano (2010): his gaze is still averted, he does not have interrogative intonation, nor interrogative morphosyntax, nor is the topic asymmetrically in his listener's purview (quite the opposite, in fact). But if a gesture hold in the absence of listener-directed gaze is a signal that impedes listener contributions (see Duncan 1972, 1973), then the relinquishing of such a hold could help to mobilize a response because it demonstrates a complete finishing of the speaker's utterance: retraction phases are a normal part of the process of finishing utterances, along with intonation and lexico-syntactic factors. Any one of these can suffice in the absence of the others, because human cognitive ability includes the principle of metonymy.[5] Thus, under the right circumstances, gestural factors besides gaze may be as influential as the four response-mobilizing factors emphasized by Stivers and Rossano (2010) and the utterance-completing lexico-syntax emphasized by De Ruiter et al. (2006).

Example 2.2 also exhibits more movement than in most cases of what I refer to as 'holds': the gesturing limbs are waggling up and down rather more than they do in most other cases, and the speaker also exhibits an exceptional amount of repetitive, abortive restarts in his speech. Unlike Example 2.1, there is no span in which his whole body goes into stasis, meaning that such processes are not a mandatory part of disfluency even if they are frequent. Broadly speaking, the whole episode in Example 2.2 constitutes a long, terminal hiatus, but we could also mark out each individual restart, and attempt to treat each cycle of his arms as a brand new gesture that repeats, much like his speech, almost the same content as the previous cycle. For my purposes there is no need to draw such fine borders; the point here is to show that during a disfluent episode, with both speech and gesture affected, the basic imagery of gesture (much of which would have been evident here even if the gesture were completely stationary) can supply some of the information missing from speech. In some examples, including this one, the extended span of the hiatus may be what allows the nearly arrested gesture to display more than

[5] In its most general sense, metonymy involves the ability to recognize whole domains of experience from partial input consisting of individual elements of those domains, and vice versa. It is among the most fundamental of human cognitive abilities; coupled with social interaction within culture, it is one of the necessary building blocks of constructs such as the *frame* (Goffman 1974; Fillmore 1985). Phenomena deriving ultimately from metonymy, among other cognitive mappings within and between domains, are the primary source material of the field of cognitive linguistics (see Sweetser and Fauconnier 1996; Lakoff 1987; Fauconnier 1994, 1997; Dancygier and Sweetser 2005).

purely stationary structure: by waggling, rather than freezing, the gesture in Example 2.2 is able to display a vague size range rather than a precise one.

It could also perhaps be labeled a "Butterworth" in McNeill's (1992, p. 77) sense, because it is being performed during a speech failure while the speaker is at a loss for words, and indeed it begins *after* the disfluency has already begun (unlike the example in Chapter 1, and unlike many other examples presented in this chapter). McNeill coined the term, somewhat humorously, in honor of Brian Butterworth's insistence that gestures occur in response to speech failures (Butterworth and Beattie 1978; Butterworth and Hadar 1989). However, the problem with the "Butterworth" category is that it seems to draw a sharp line between ordinary co-speech gestures and those purportedly very rare gestures specifically *caused* by speech failures and performed only while trying to recall a word.[6] Even though I argue that a gesture's duration could possibly be extended by co-timing with speech that includes a hiatus, this does not mean the gesture itself is caused by the speech failure and performed expressly for the purpose of aiding lexical retrieval. It may certainly affect lexical retrieval, but this is also true of everything else affecting a speaker's attention and thought processes: the immediate linguistic context is of paramount importance for lexical retrieval, yet surely we do not claim to produce words mainly for the purposes of lexical retrieval of subsequent words. As McCullough (2005) suggests, lexical facilitation can be part of our model without the need to downplay communicative function, as some authors have (cf. Krauss, Morrel-Samuels, and Colasante 1991). We can acknowledge an extended gesture's impact on concurrent and subsequent semiosis by describing how it participates in ongoing contextualization, whether transformative or as maintenance, in concert with all the other elements of a multimodally produced utterance. The speaker cannot escape the impact of this visuospatial contextualization, yet neither can an interlocutor.[7] Communicative function and lexical facilitation effects could easily be inseparable aspects of a single process.

Examples 2.3 and 2.4 — gesture and gaze facilitating completion by the listener

The next two examples demonstrate that, in addition to providing crucial information about the intended utterance, gestures persisting during disfluency are sometimes used by listeners to offer suggestions for completion of a speaker's absent linguistic stream. In

[6] Duncan (1996, p. 86) mentions an intriguing possibility relating to her own examples of progressive aspect appearing as repeated cycles of movement in gesture. She notes the existence of repetitive "Butterworth beats" coinciding with lexical search, in which the sustained repetitive cycle may metapragmatically display the speaker's ongoing process of lexical difficulty and help indicate an intent to continue speaking. In Chapter 1, I suggested that by *bracketing* a span of time with a sustained gesture configuration, a speaker may be displaying (not necessarily on purpose) that they are 'doing' some topic. This mechanic could also apply across disfluent spans, perhaps as an automatic embodied result of unplanned hiatus. If then combined with an overlaid repetitive motion further emphasizing the progressive aspect, such behavior during a speech hiatus could be an even stronger display of 'doing' disfluency.

[7] Streeck (1993, pp. 296-297) provides an excellent discussion of these issues: speakers and listeners are influenced by their own gestures, just as they are influenced by any other external artifacts brought into play. This is an important theme in Chapter 4 below.

contrast to Example 2.2 above, in which the speaker did not gaze at his interlocutor during his hiatus and she accordingly did not offer a suggested completion, in both of the following examples the speaker is gazing intently at the listener during the hiatus, and the listener 'chimes in' to fill the gap in speech.[8] It should be noted that the presence or absence of listener-directed gaze does not always determine whether a listener will offer a contribution during a speaker's hiatus, in part because cultural expectations regarding listener feedback are highly variable. For example, Hayashi (2005, p. 28) discusses a case in which the listener suggests an interpretation after the original speaker has withdrawn her gaze and is directing it at her gesturing hands.

Example 2.3 below contains two distinct instances of a gesture hold persisting across a hiatus in speech. The first case very subtly presages future specification in speech, while the second case coincides with listener-directed gaze and presages the rest of the utterance powerfully enough to allow the listener to suggest a successful completion. First of all, in images [b2] and [b3] we see that Participant A has extended the thumb, index finger, and middle finger of each hand, indicating 'three' entities. This hold lasts for more than 2 seconds, through an additional stroke which finally synchronizes with "three" (image [b3]) and makes clear what the 'three-ness' of her gesture is connected to. During the period shown through image [b3], her gaze is directed only at her hands and the model theater, and not at her interlocutor, who remains motionless while alternating her gaze between her partner's face and the held gesture. It seems likely that Participant A's thought processes may be affected by the long-lived 'three entities' hold, but it is difficult to say whether Participant B's uptake is affected by having visual access to it.

In contrast, between images [b3] and [c1], Participant A returns her gaze to her partner's face while raising and holding her hands in the configuration shown in image [c1]. At the end of this speech hiatus (filled first with silence and then with "kind of") she thrusts the held configuration down in a stroke that is not synchronized with any speech of her own. However, her partner immediately fills in the gap with a suggested completion (during which Participant A's right hand actually beats down slightly). Unlike with the 'handoff' gesture phenomenon I will discuss in Chapter 5, Participant A does not drop her hold and release the floor to coincide with her partner's contribution. Instead, she simply continues her utterance (with a lexical choice that may be a harmonization to her partner's). Her hands pulse with another beat on her repetition of "corner," followed by a new phrase of fluent speech and gesture to finish the utterance.

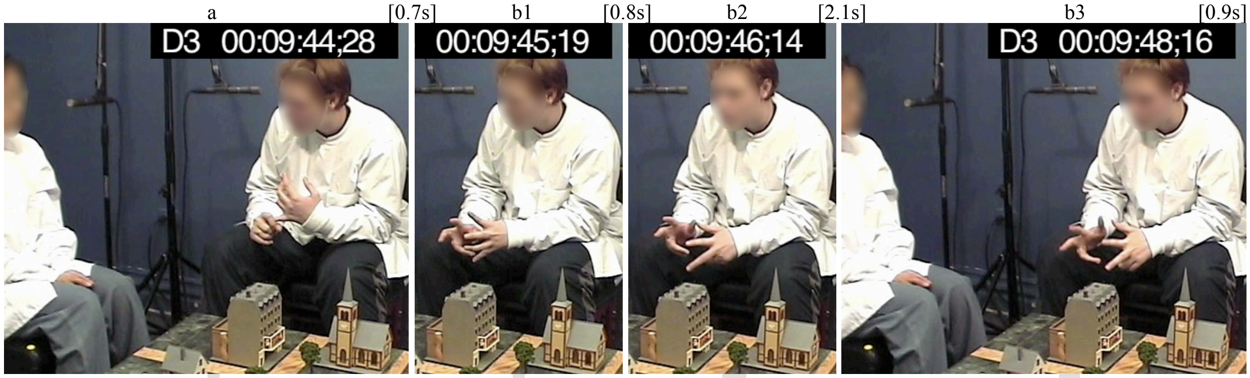
It is interesting to note that she drops to rest at the beginning, rather than the end, of her final pulse of speech (see images [d] to [e]), even though this is the lexical material which could be interpreted as most closely related to her preceding gesture. Examples

[8] These examples share a great deal in common with an example of Streeck's (1993, pp. 292-293, 1994, pp. 252-255). It should be noted that in the earlier paper, Streeck emphasized the importance of the speaker's gaze onto her own gesture, a behavior which has not usually been specifically associated with the mobilization of immediate listener response (though it does increase the likelihood of listener recall of gestured content: see Gullberg and Kita 2009). Crucially, Streeck later modified his description to state that just before the listener's contribution, the speaker gazes toward her "and, thus, invites her to join" (Streeck 1994, p. 253).

like this have led some authors, notably Schegloff (1984), to make the strong and oft-repeated claim that gestures tend to greatly precede their associated "lexical affiliates." But the end of this example once again illustrates that the initial gesture and speech are quite synchronized: the idea is 'force them out the back', captured holistically in gesture even though the synchronized speech includes only two lexemes. The additional speech, once added, fully unpacks this idea linguistically, and this coincides with a gesture retraction, intonation, and speech rate that 'let go' of the utterance and signal its closing.

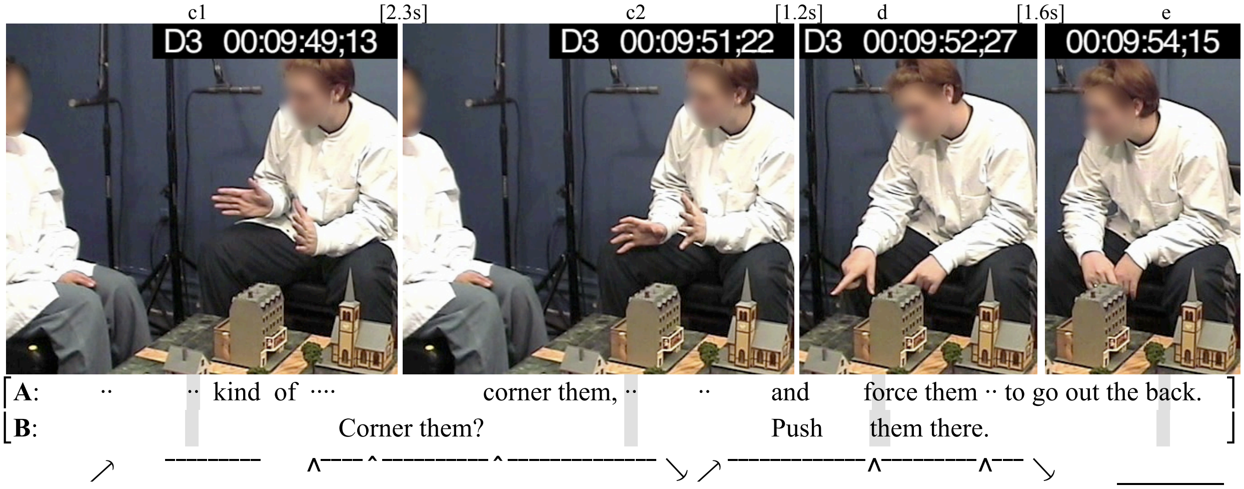
(2.3) **B**: Now, ·· they could just keep evading us· though and not · ever ·· go out the doors.
What would incite them ···· to go out the doors? |**A**: Hmm ··
A: Well maybe·· · when you get in there·· you want to block off* ·····
Like · uh·· you know how you c* · lobby* · the doors to the lobby? ·
So that they can't get out that way, #
And that way if you guys know ex[a]actly where they're loc[b1]ated? · #
you [b2]should be able to ···· with thr[b3]ee of you
[**A**: [c1]·· kind of ···· corner them, [c2]·· ·· and f[d]orce them ·· to go out the b[e]ack.]
[**B**: Corner them? Push them there.]

a-b1: *From an 'enumerative' hold with RH fist grasping LH pinky, she switches to BH symmetrical indistinct claw shape possibly indicating the location of multiple entities in an abstract space.*
b1-b2: *This hold changes slightly with a small downward stroke, such that BH now have thumb, index finger, and middle finger extended while other fingers remain curled.*
b2-b3: *This held config. lasts through a long pause as she looks at the model. The final stroke, timed with "three," takes place 2 seconds after the earlier stroke where this configuration appeared.*



A: **(…)** know exactly where they're located? · # you should be able to ···· with three of you b3-c2: *BH rise and switch to palm-out closed 5s with the palm heels forming a right angle, which is held in place (with gaze on her partner) through her filled pause ("kind of"), before a small up-down stroke at* [49;29] *as she lapses into silence, with gaze still directed at her partner. Participant B responds by suggesting a sentence completion, which times with a small downward beat from Participant A's RH. Participant A repeats the verb phrase (timing with another small beat from BH), and the hold as well as her partner-directed gaze persists through a following pause.*

c2-e: *She drops the 'corner' hold as well as her partner-directed gaze, and switches to a dual index finger deictic hold at side of theater, which she sweeps toward its rear (with the stroke falling on "force") and holds in place, then returns and sweeps again (with the stroke timed with her momentary pause), and then drops to rest as she begins the final part of her utterance.*



A: ·· ·· kind of ···· corner them, ·· ·· and force them ·· to go out the back.

B: Corner them? Push them there.

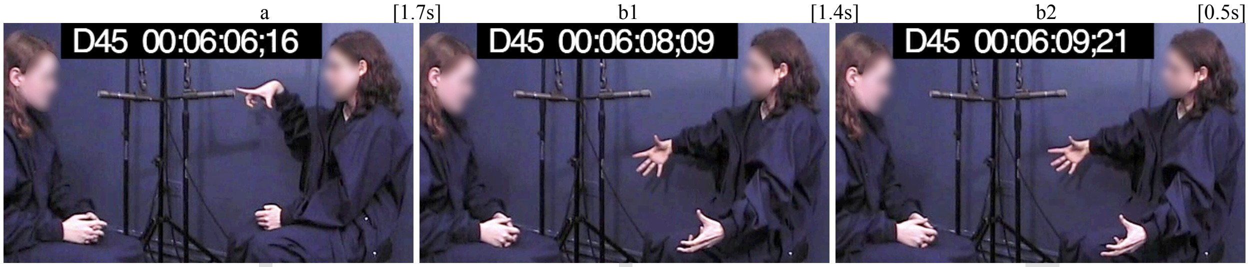
Another case of listener completion during hiatus is illustrated by Example 2.4 below. In this example, Participant A begins a 'hefting' gesture at the very end of the speech preceding her lengthy, mostly silent hiatus. Participant B remains motionless and silent until, as shown in image [b3], Participant A lifts her gesture to a higher position and looks up into her partner's eyes, at which point Participant B offers a completion. This time, unlike in Example 2.3 above, Participant A allows her partner's suggestion to stand as the utterance's 'official' completion, drawing in her hands as she assents.

Examples 2.3 and 2.4 share important characteristics: in both cases, the listener remains silent and attentive during the first part of the speaker's speech hiatus. Then, seemingly in response to a gaze trigger, the listener offers linguistic material to fill in the gap left by the speaker's hiatus, even though the original speaker could easily be intending to continue the utterance. In both cases, although the listener could have conceivably come up with the suggested completion without being influenced at all by Participant A's gestures, it seems far more likely that gesture played a major role in broadcasting the total intended meaning of Participant A's incomplete utterance. Indeed, we know that people glean information from co-speech gesture that they could not get from speech alone (Beattie and Shovelton 1999a, 1999b, 2001; see also Kendon 2004, ch. 10). Therefore, to claim that Participant B in Examples 2.3 and 2.4 was somehow not influenced by gesture, we would also need to claim that people can selectively avoid observing informative cues placed saliently in their visual fields. Given that the listener in these examples was already engaged in the task of attentively trying to learn her interlocutor's meaning, this seems quite implausible here.

```
                                                  a                           b1      b2
(2.4)   A:  The tarp can't be so big · that ·· each person· can't ····· uhm

             b3               b4              c
       ⌈A:  ···                            Right. · ⌉
       ⌊B:      manage it by themselves.            ⌋

                  d1                 d2          e1      e2
       ⌈A:  Like once the wombat steps on it, ·· they're gonna have to pull it up I guess.  ⌉
       ⌊B:       Yeah.                                                              Yeah. ⌋
```

a: *RH has been holding in an abstract 'discourse deictic' with thumb and index finger extended.*

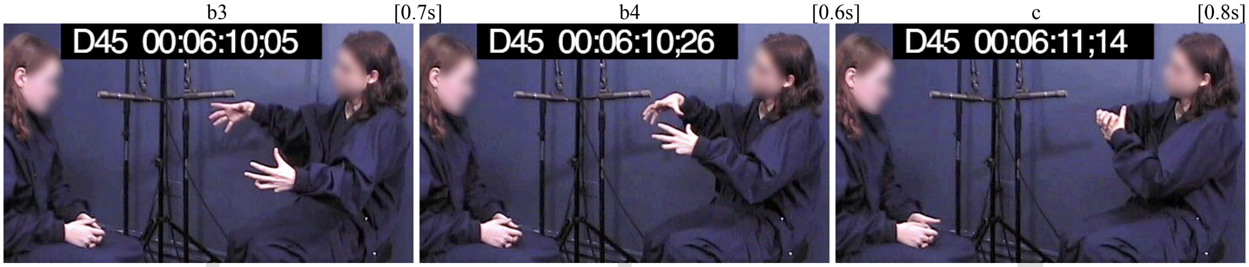
a-b2: *BH then switch to a symmetrical, 2-foot wide 'hefting' gesture which she holds with a slow up-and-down oscillation, through 2 seconds of pausing, with gaze directed at hands.*



```
A:  (...)  be so big · that ··    each    person·    can't  ·····--------------------------uhm----------
(A's RH) ---^-----------------------------  (BH) ↗   ∧  ------------------------------------------------
```

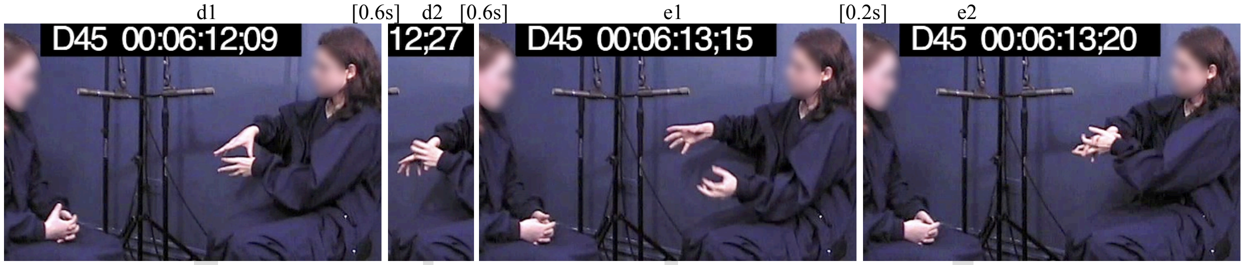
b3: *She raises her arms and orients her head and eyes up to her partner, and continues 'hefting'.*

b3-c: *Once her partner begins suggesting a sentence completion, she draws in her hands and eventually clasps them together at momentary rest, in time with her verbal assent.*



```
⌈A:              ···                                            Right.      ·  ⌉
⌊B:         -----------manage--------it--------by--------them---- selves.      ⌋
(A's BH) ↗  ---------------------------------------------------- ↘ ________ ↗
```

c-d1: *BH then thrust down and right on "once," presaging the wombat's stepping motion onto the tarp.*
d2: *This motion is reiterated via a LH stroke timed with the verb "steps" itself as RH stays in place.*
e1-e2: *BH then rise and clasp together forcefully, showing the action of 'pull it up', which becomes fully specified in speech nearly 1 second later.*



A: Like once the wombat steps on it, ·· they 're gonna have to pull it **(...)**
(A's RH) ---↗ ∧ ———————————————— ↗ ∧ ————————————
(A's LH) ---↗ ∧ —————————— ↗∧ ——— ↘ ↗ ∧ ————————————

At the very end of Example 2.4 above, the speaker's gestures coincide with fluent speech rather than any sort of hiatus, and I include the images here to again illustrate how, especially from a listener's point of view, and during fluent as well as disfluent speech, a speaker's gestures often 'presage' chunks of the utterance meaning by various degrees. Although the gesture is almost always synchronized with *something* appropriate, it is worth investigating whether a listener's uptake is influenced by the small interval before all of the explanatory (or 'co-contextualizing') linguistic items are spoken. For example, image [d1] shows that Participant B is already observing something like a stepping action performed by Participant A's hands, synchronized with "once" and occurring 0.6 seconds before the verb "steps" is spoken. The observation of this enactment will unavoidably affect her constantly updating beliefs about her partner's intended meanings. We need to investigate how it can create an imagistic 'pre-understanding' that could influence the efficiency of her uptake of the rest of the spoken utterance. Similarly, images [e1] and [e2] show another action synchronized with "they're" (and then "they're gonna" and then "they're gonna have to," etc.). The subject of the action ("they're") is already in play, coupled with an enactment in gesture of what "they" will be doing.

This is a perfectly logical point of synchronization, fully consistent with McNeill's Growth Point Theory, and this is not negated by the fact that the action is further specified by later speech. But from an utterance comprehension standpoint, we should investigate whether Participant B's uptake of the verb phrase "pull it up" is affected by having viewed the full action's enactment *nearly one second earlier.*[9] In fact, she may be

[9] Indeed, Kelly et al. (2004) and Wu and Coulson (2007) found neurocognitive evidence that gestures observed approximately one to two seconds before speech had an effect on the comprehension process, depending on the degree of semantic mismatch between speech and gesture. Özyürek et al. (2007) and Willems et al. (2007) found similar effects when speech and gesture were fully synchronized. These studies lend plausibility to my claims that listeners' comprehension processes are affected by access to complementary information from gesture, as they hear words in a stream of speech.

expecting this very verb phrase (or something quite similar) already, as evidenced by the listener responses during other speaker disfluencies in Examples 2.3 and 2.4. These ‘presaging’ gestures will have a profound effect on utterance comprehension (or, more appropriately, they likely *are* a profound *aspect of* utterance comprehension), since they can reveal strong clues about what a person will say right from an utterance’s inception.[10] It also seems likely to be a completely normal aspect of *all* normal face-to-face conversation, worldwide: there may be many local variations, but there are no human societies which lack a time-spanning process of linguistic segmentation or the ability to enact basic actions and iconic representations in gesture.

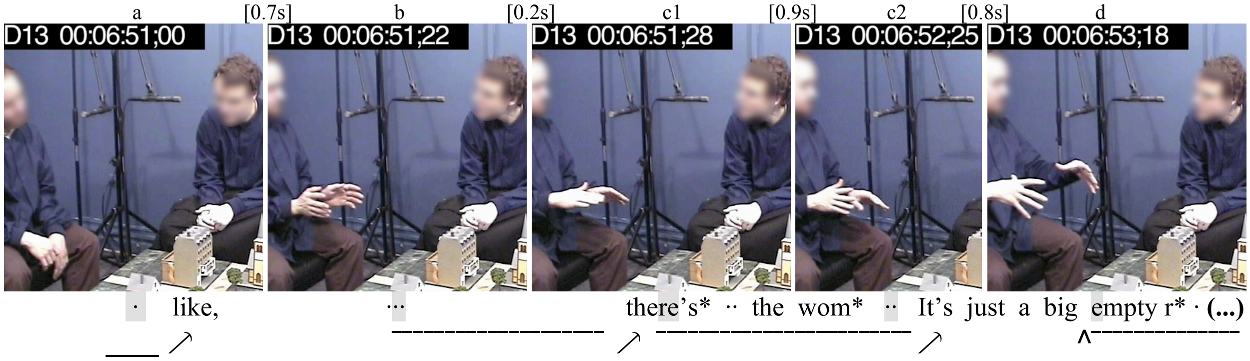
Example 2.5 — a gesture whose context reduces its inferential power

Finally, the last example in this section serves to illustrate how much the interpretability of a hiatus-spanning hold depends on its immediate context. Most of the gestures in the preceding examples were easily interpretable as stand-alone contributions to the utterance (an exception was the hold representing ‘three’ in Example 2.3), even though they could also have served as co-speech gestures. Depending on context, the hold in Example 2.5 below could also be readily interpretable, but in this case the linguistic context effectively cancels this possibility. The evolving hold begins on the word “like,” and eventually it becomes clear that it represents the boundaries of a basic indoor space. However, unlike the size-delimiting gesture in Example 2.2 above, the disfluent speech in the example below instead seems to muddle the referential power of the gesture. The discourse leading up to the passage shown in the images helps build a strong expectation that Participant B’s utterance will be about the details of the building under discussion. In fact, this turns out to be correct. But during the hold shown in images [c1] and [c2], this assumption on the part of the listener may be disrupted temporarily, because of the unexpected mention of “the wom*(bats).” It will then be restored once Participant A starts speaking more coherently again and revises his gesture to be more in the shape of an auditorium, as shown in image [d]. Thus, there is a great deal of variation in how effectively a listener can be expected to guess at a disfluent utterance’s trajectory, depending on how interpretable a hiatus-spanning gesture is in context.

[10] When this process occurs via gesture, it is usually revealing specific referential details, which means it is triggering the “particularized conversational implicatures” (or PCIs) mentioned by Levinson (2000). Levinson’s work is devoted primarily to the related phenomena of “generalized conversational implicatures” (or GCIs), which involve the effect that certain lexico-syntactic choices at the start of an utterance, such as quantifiers, have on the likely meaning of the rest of the utterance. Gesture may well be able to activate GCIs as well. Regardless, the important features of the temporal model are the same for PCIs and GCIs: pragmatics (i.e., ‘context’) is viewed as always-on, constantly influential and evolving, and always taking scope over any process of semantic decoding, including that which a listener must conduct as new words are added to an utterance. Unless linguistic models of utterance comprehension include the influence of gesture and other non-linguistic cues, they will continue to ignore an incredibly rich amount of scope-defining information observed by the listener ahead of the moment of semantic decoding. Although Levinson (2000) does not discuss gesture, the architecture of his model leads directly to the conclusion that a pragmatics without gesture is woefully incomplete.

(2.5) **A**: # ····· Yeah, but I think* · I think it's a movie theater. It's not a hotel. ·· So.
B: Oh a movie theater. ·· Yeah but like ·· so what's all these floors about, like ··
A: # It's a movie theater. You need to have like, y'know.
Y'know how you have like the balcony and then you have · the ····
B: I guess that's true.
A: orchestra.
B: # So you think there's like, #
a b c1 c2 d
· like, ··· there's* ·· the wom* ·· It's just a big empty r* · auditorium-style room?
with like · the screen?
or # like there's actually like lots of levels so the wombat to like sneak up into

a-c1: *BH rise from rest on the word "like" and hold with palm heels facing each other for less than half a second. With the return of speech, the wrists rotate so palms face downward.*
c1-c2: *BH hold in place unsteadily through two self-interruptions of seemingly unrelated speech.*
c2-d: *As relatively fluent speech returns, BH spread their fingers, rise several inches, and thrust down in a small stroke, outlining a more well-defined bubble of space, which is held for 1.5 seconds.*



These examples have shown that the co-emergent pulses of speech and gesture envisioned by McNeill (1992, 2005) are frequently fragmentary and incomplete, in such a way that the often holistic imagery of gesture can become extremely informative in the event of a breakdown in the spoken linearization of co-emergent speech. The current Growth Point model may be too dedicated to discrete, self-contained and complete packages of speech and gesture. McCullough (2005) has emphasized that it needs to be elaborated to account for longer, cross-clausal intervals of sustained contextualization, but it also needs to allow for the tying together of brief and fragmentary 'abortive' Growth Points which nonetheless can be quite functional in conversation.

The examples in this section and in the rest of the dissertation raise problems for our definitions of a contribution or utterance "move" composed of a composite of multimodal elements (Enfield 2009, Kendon 2004). The edges of the "move" are often unified across multiple modalities, but at other times, one stream of content may appear to abort or falter while another stream continues to 'deliver'. This can happen even in the absence of intention to do so on the part of the speaker, because of such mundane facts of reality as the inertia of physical matter and its stronger hold on the body's most massive articulators, compared to its effect on the articulators of the vocal tract. The examples discussed so far have shown that a move may still be successful even when seemingly crucial parts of it are missing. I now turn to cases where, although the elements of the move appear to have reached a natural and successful conclusion, the move nonetheless persists.

Holds maintaining retrospective links across pauses and asides

As in the cases shown above, the following examples all involve the maintenance of a gesture hold across a hiatus in speech. They also differ strikingly, however, because the holds in the cases below originate from a prior moment of synchronized speech and gesture which is then followed by a post-stroke hold that spans the hiatus. As a result, there is no ambiguity about the meaning of the speech-gesture combination in context. Rather than lasting through the hiatus as a de facto proxy for a not-yet-realized speech-gesture composite, the hold instead remains as an *artifact of that successful combination.*

Examples 2.6 and 2.7 — maintenance of retrospective imagery across a long pause

In the first of this pair of examples, the speaker twists her wrists to enact a 'locking' motion twice, first with the verb ("lock") and then with the object ("doors"). After the second stroke, she goes into a long stasis in both gesture and speech (note the large amount of time between the final three images), finally filling it with "uhm" as her muscles very slowly relax, allowing her gripped fingers to loosen and her wrists to untwist somewhat. Only after an additional long silence does she suddenly break out of her muscular stasis at the same time as resuming fluent, rapid speech. Throughout this period, she keeps her gaze fixated on her interlocutor, but achieves no response.

In some ways, Example 2.6 is like a long-form version of Example 2.1, but this time, the form of the gesture is referentially significant because it retains the shape used in the 'locking' gesture. More than two seconds after the stroke, much of this shape remains as an artifact available to the perception of both participants (for the listener, it is available near the center of her visual field, while for the speaker, it is peripherally visible but mainly available to her kinesically). This retrospective artifact may help bridge the gap between her preceding fluent speech and that which follows. It could also be analyzed as *pro*spective, however, because its continued embodied presence means that the speaker will have to begin with it as her starting point, when she resumes speaking. This forward-linking effect is bolstered by the incompleteness entailed by slightly rising intonation on "lock" and "doors," meanwhile the linguistic string itself is consistent with either continuation or termination. Reversing perspectives, we could similarly say that the continuation implied by her intonation is bolstered by the maintenance of the gestural artifact during the hiatus.

(2.6) **A**: So maybe, ··· uuhh ···· after we get the cages set up, when you guys go in,

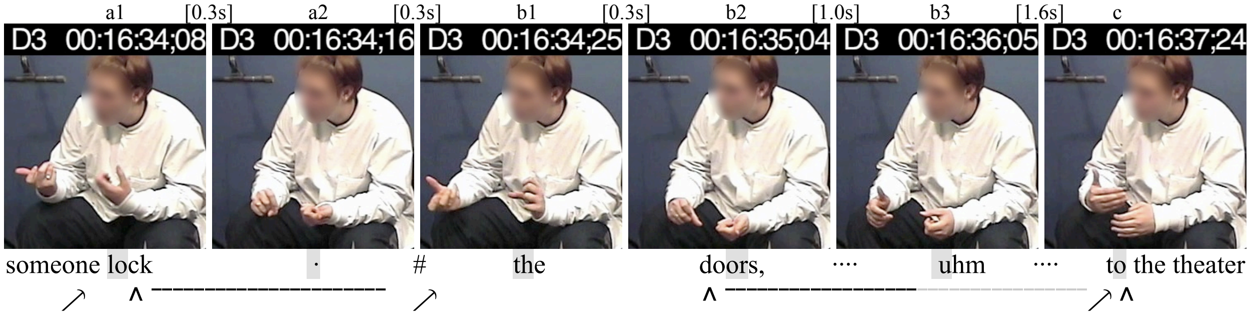
a1 a2 b1 b2 b3 c

we'll have someone lock · # the doors, ···· uhm ···· to the theater from the lobby

a1-a2: *While gazing at the listener, BH have risen from thighs into loose gripping shapes, then twist together at the wrists toward her left (RH twists a full 180°) simultaneous with a downward stroke during which the hands become gripped firmly. This configuration holds in place momentarily.*

b1-b3: *BH repeat this motion (this time, LH wrist turns over during the preparation, and thus is able to nearly match RH's full twist), after which they hold in place as her speech enters a hiatus. During the more than two second pause, her gripped hands relax somewhat and her wrists slowly untwist by about 90°, but they remain in place and retain most of their shape.*

c: *Simultaneous with the return of fluent speech, her palms flatten and flap towards her and away in a set of rapid gestures which contrast strongly with the preceding held configuration. On the second syllable of "theater" she finally averts her gaze away from the listener momentarily.*



Example 2.7 below has similar characteristics: the linguistic string before the hiatus could, in theory, form a complete utterance, however the speaker's non-final intonation implies otherwise, and her gestural hold persisting into the hiatus is similarly "non-final." During this silence she even repeats the gesture stroke and hold, re-emphasizing the retrospective link to the previously stated idea of "aisles on the side." By repeating this gesture during the hiatus, she also increases the likelihood that the idea is meant to contextualize whatever is to follow. As long as the gesturally maintained idea is kept 'at hand', it is the topic going forward. Her gaze remains fixated on her partner throughout the entire exchange, so there is no gaze "trigger" for her partner to respond to, but Participant B does offer the small backchannel "okay" simultaneous with Participant A's "uhm," just after the latter lets her arms sag slightly (image [c2] to [c3]). Dropping of the hold does not necessarily signal a closing of the topic, but it does signal the end of Participant A's asymmetrical control of the floor. Just after the portion shown in the images, she again lapses into a speech hiatus (with gaze still fixated on her partner), but this time her hands are at rest and Participant B finally offers a full contribution.

(2.7) **A**: If ·· my assumptions are correct

[**A**: there's just one sec[a]tion of seating and [b1]aisles on the si[b2]de? ·· [b3]·· [c1]·· [c2]·· # uhm [c3]···]
[**B**: Okay.]

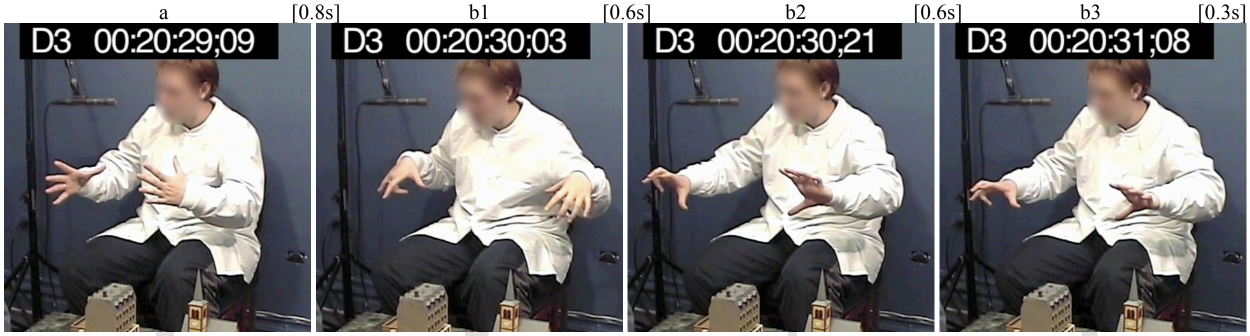
A: So that [d]might make it a little more ····

B: Oh, ·· that makes it a little different doesn't it.

a: *After fixating her gaze on her partner's face with the word "one," BH (spread 5s) have thrust forward and are holding around a 1-foot-wide 'seating' area. Her gaze remains throughout.*

a-b1: *She retracts her arms to the sides of her body in a wider stance with fingers vs. thumb of each hand preparing to outline the boundaries of 'aisles' on each side.*

b1-b3: *BH thrust forward to outline the 'aisles', and this position is held in place for nearly 1 second as she lapses into a silent speech hiatus.*

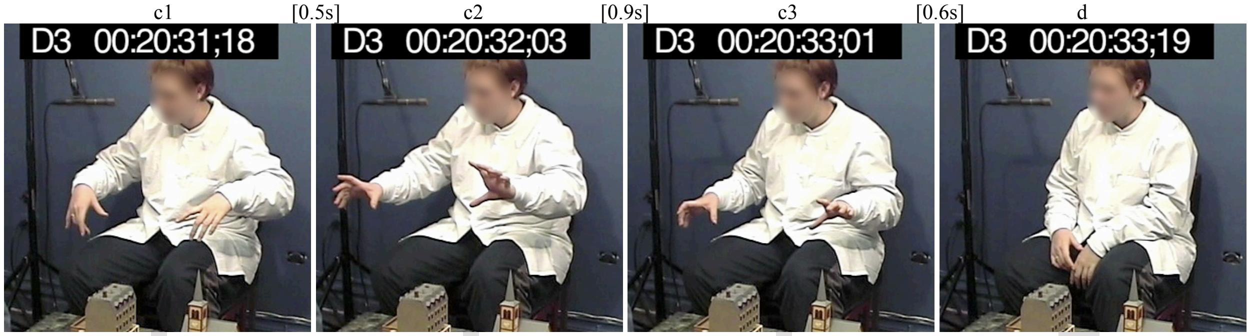


just one section of seating and aisles on the side? ·· ··
↗------- ^ ------------- ↗ ^ ------------------------------------

b3-c1: *Her arms remain stationary but she drops her hands at the wrist in preparation for a new stroke.*

c1-c3: *The previous 'aisles' stroke is repeated in silence, and held place again. As she takes a breath and fills her hiatus with "uhm," her arms relax slightly while her hands retain their shape. Simultaneous with "uhm," her partner nods once vigorously while saying "Okay" (not shown).*

c3-d: *BH drop fully to rest just as she resumes fluent speech. Gaze remains fixated on her partner.*



·· ·· # uhm ··· So that might make it
↗ ^ ---------------------------- ↘ ________________

The holds in Examples 2.6 and 2.7 are retrospective in that they originate as representational gestures synchronized with preceding speech content, but by being held through the disfluency they also make clear the speaker's intent to continue speaking *and* they entail that the continuation will use the held configuration as a basis for contextualization (that is, the new content will be figural against the ground of the maintained hold). Such forward-contextualizing, retrospective holds can be continued indefinitely, even as other articulators perform new gestures, as Examples 2.9 and 2.10 will show at the end of the chapter.

Example 2.8 — retrospectively established gesture content, employed prospectively

The last example in this section illustrates a combination of retrospective and prospective effects, via a series of deictic holds centered on "House 33" in the model village. Until now, almost none of the examples in this chapter have involved gestures that directly engage with the permanent artifact of the model village (and in Example 2.4 the model was simply absent), so it is noteworthy that the referents being pinpointed deictically in Example 2.8 remain accessible whether or not a gesture hold is maintained. However, with such a large variety of possible referents, the mere presence of the model does not serve to draw attention to specific items, whereas deictic gestures can serve to pinpoint particular referents and maintain their special relevance to current utterances.

The deictic hold in images [a1] to [a2] maintains a very strong retrospective link to the referent first pinpointed as "this house?" then elaborated as "thirty-three" and "the yellow one." After 4 seconds, this link is still active and can be assumed to remain the basis of what will come next. The link is 'released', somewhat, when she disengages the deictic hold (images [b1] to [b2]) while still continuing the extended hiatus from speech. However, the rest of her body, including her gaze, remains configured as before, and the hold is actually continued, with no change in handshape, though it is now pressed against her mouth. The handshape remains available to resume its former purpose, which perhaps helps her signal her intent to continue the previous topic, while also showing a temporary withdrawal as she searches her memory of the instructions. She then resumes her earlier deictic hold (images [c1] to [c2]), and maintains this link through a *fluent* hiatus in the form of a side comment affirming the 'good enough' status of the narrative content she has presented about the deictically pinpointed referent. Unlike the disfluencies discussed, this fluent aside is a controlled self-interruption. The utility of the hold, however, remains the same: it extends an artifact of a referent beyond the moment it is first focalized in speech and gesture. In Chapter 4 this same utility is shown to apply, in exactly the same way, across interruptions caused by others.

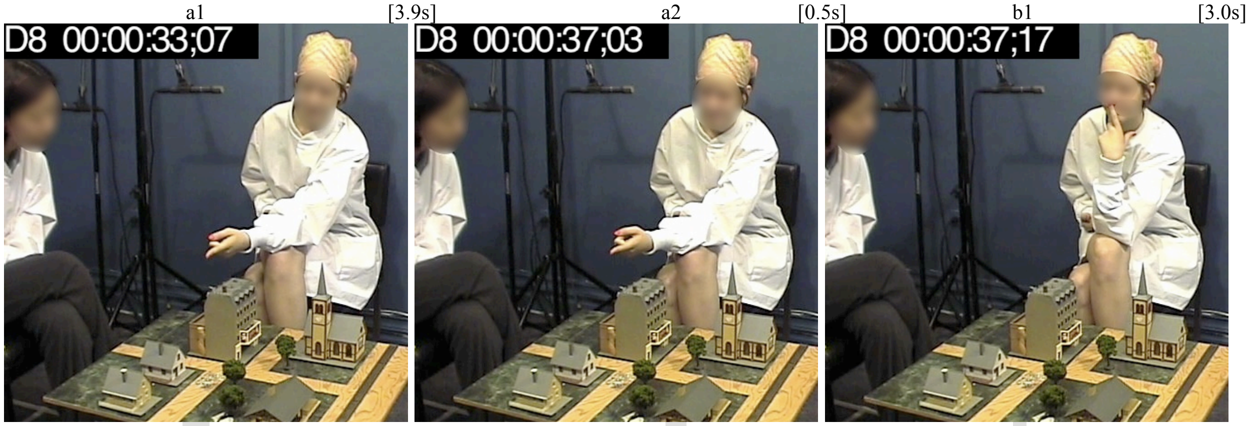
Then, at the moment shown in image [d1], an interesting shift occurs: she alters her posture, sucks in her breath, turns over her wrist and extends her arm, all of which are clear precursors to a an utterance which would resume the narrative description regarding House 33. This resumption is interrupted by a burst of laughter, and thus the remaining hold maintains *prospective* reference to the not-yet-realized new utterance, though the hold's referent is also clearly established at this point and so it maintains retrospective links to preceding discourse relating to it. During her phatically infectious laughter and temporary gaze shift to her interlocutor (who offers an incredulous backchannel response), Participant A's muscles relax somewhat (images [d2] to [d3]) but the deictic hold is not canceled, and she resumes narrative focus on the referent with another slight twist of the wrist (image [e]), at which point she finally begins the spoken part of the long-expected resumption of the task description.

```
(2.8)  A:  And then, ···· we have to go,* ····· Oh, shoot.

                                        a1
           ·· We have to go to this house? · We have to go to thirty-three first. The yellow one. ··

                                 a2      b1
           {tongue-click}  ····  Uhm   ············  I don't r* ··

                    b2          c1                                    c2
           Maybe I pick you up there I don't remember. ·· Yeah. ·· Something like that. ··

                 d1                            d2                    d3
       [ A:  #   ··   {tongue-click}   @@@@  #           @  @  #    ]
       [ B:      #                                Okay? @@   #  @    ]

                                  e                               f
       A:  But then from house thirty-three we have to go·· to·· this one house thirty-five
```

a1: *Gaze remains fixated on the model, and LH was at rest draped over right knee (not shown). LH has now risen to point index finger, palm-up, toward the house at the far lower-left of the image.*

a1-a2: *The hold is maintained unchanged for 4 seconds, first through 2 seconds of further speech (see transcript above), then through a tongue-click and pause of nearly 2 additional seconds, before she breaks the silence with "Uhm" just before altering the gesture.*

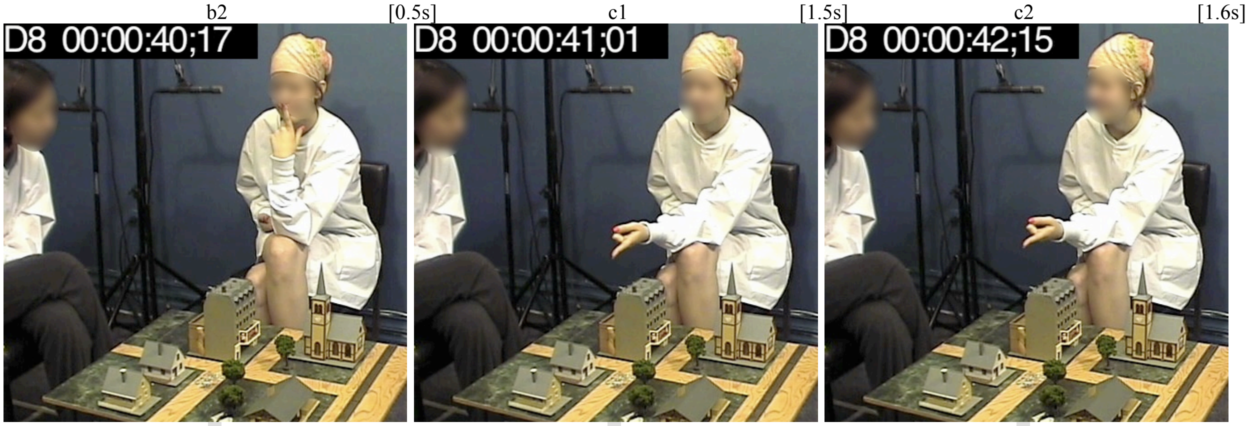
a2-b1: *LH rises without altering its handshape, to rest against her mouth for 2 more seconds of silence followed by nearly 1 second of disfluent speech.*



```
(...)  go to this house? · We have to (...) {tongue-click} ···· Uhm          ···                ········ I  don't  r* ··
 ↗                  ^ ————————————————————————————————————————————    ↗                       —————————————————
```

b2-c1: *Upon the return of fluent speech, LH resumes its deictic hold, with the same handshape and orientation as before, on House 33 at the lower-left of the image.*

c1-c2: *During the hold she turns head and gaze momentarily toward her interlocutor (on "yeah"), before a brief metanarrative comment ("something like that") while returning her gaze to the model.*



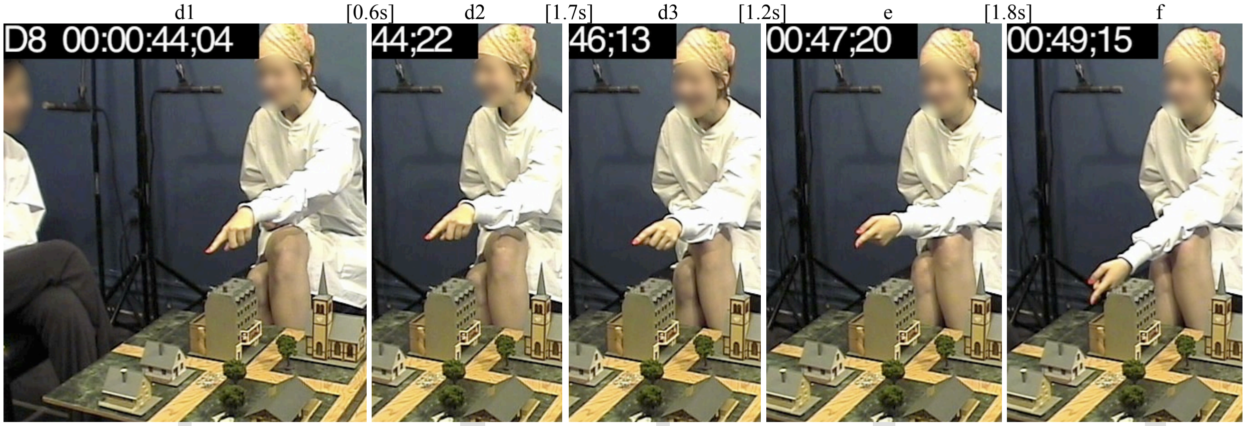
Maybe I pick you up there I don't remember. ·· Yeah. ·· Something like that
——————————↗ ——————————————————————————↗

c2-d1: *She wordlessly marks resumption of the narrative by re-establishing strong deictic focus on the house. This is accomplished via a slight raising of posture and a further extension of her arm that includes a 180° twist of the wrist, simultaneous with a loud nasal intake of breath.*

d1-d3: *The narrative resumption is then interrupted by 2 seconds of laughter, during which the hold is maintained but loses some muscle tension, as her gaze goes to her partner then back to the model.*

d3-e: *Strong deictic focus is once again reinstated along with fluent speech about House 33.*

e-f: *While maintaining the same handshape, her arm lowers to indicate the neighboring house.*



# ·· {*tongue-click*} @@@@ # @ @ # But then from house 33 we have to go·· to·· this one
^ ——————————————————————————^——^—————— ↗ ^———^———^———

During this example the deictic hold seems likely to support *both* the speaker's efforts at maintaining narrative coherence through various pauses and memory searches, and the listener's own ability to track the semantic scope of the speaker's utterances, even when one is disrupted by laughter before it has begun. I believe the almost certain effect on Participant B should be clear and convincing to the reader, in the role of distantly removed proxy interlocutor, even through the impoverished medium of these screenshots. In context, the moment shown in image [d1] broadcasts not just an impending move but also a rather narrow scope for its content. Even with a total breakdown of speech before the first word is spoken, the resumed deictic and its maintenance through the interruption are a veritable flashing neon sign of Participant A's 'in-order-to motive'—she *will* resume narrative discussion about House 33. The listener will have become fully oriented to that specific scope, long before the move is successfully begun.

Holds persisting during new gestures by the other hand: The 'buoy' phenomenon

The final examples in this chapter deal with a striking phenomenon which shows clearly the ability of a speaker to leave a gesture hold in place while performing other gestures. This would not be possible without the particular affordances of the "kinesic medium" (Kendon 2004), in particular its independently moveable and independently observable articulators, which are part of the common "human equipment" of "every intact member of the species" (Ekman and Friesen 1969, p. 59). By way of contrast with the more limited affordances of the vocal tract's sound-making apparatus, Kendon (2004, p. 310) puts it best:

> In the kinesic medium it is possible to structure expressions spatially as well as sequentially. Further, because the bodily instruments that any kinesic code makes use of - the two hands, the head, the face, the eyes, the torso, and so forth - are spatially distributed, and because, to a degree, they can be moved differentially in relation to one another, a kind of orchestration of bodily instruments is possible. That is, more than one thing can be signified kinesically at the same time. Furthermore, it is also possible to use some bodily expressions in such a way that they frame or bracket together expressions by other parts of the body. Thus, facial gestures, such as eyebrow movements or positionings, movements of the mouth, head postures and sustainments and changes in gaze direction can, and do, serve as important means by which relationships between successive expressions, such as those produced by the hands, can be linked together in various ways.[11]

Here Kendon speaks directly to the *bracketing* capacity discussed at the end of Chapter 1. What my examples will stress, however, is that there is no hard boundary between gestures whose stroke phase is timed carefully with speech, and gestures that can "frame

[11] It is important to note that the vocal apparatus is not completely devoid of affordances allowing for independently observable tiers of signals. For example, voice quality, rate of speech, and other parameters can be modulated during speech in meaningful ways. However, all signals originate from the single region of the mouth, therefore the vocal apparatus truly lacks the kinesic medium's spatial structuring ability.

or bracket together expressions by other parts of the body." The linking factor is the post-stroke hold, which can simply be left to persist as new speech-gesture composites are performed with the rest of the body. To sustain a body configuration as the ground against which manual gestures are performed, a speaker is not limited to just facial expressions, postures, and gaze—earlier manual gestures themselves can perform this function. Kita, van Gijn, and van der Hulst (1998) required a division between post-stroke holds and 'independent holds' in their revision of Kendon's (1980) architecture of the Gesture Phrase, but this would entail the requirement that all parts of the body participate simultaneously in the division between one utterance and another. As McCullough (2005) stresses, such a requirement fails to allow the possibility that each new utterance consists partially of inherited structure carried over from previous action, and partially of new structure created via transformative changes. Crucially, a speaker may not need to consciously include such pre-existing structure to allow it to persist during new Gesture Phrases; instead, when congruent with the new phrase and when the limb in question is not required for new gesturing, persistence could be the default.

The phenomenon of a gesture being held in place with one hand while the other hand performs a new gesture is known as a 'buoy', a term inherited from Liddell (2003), who observed an analogous phenomenon in American Sign Language. Although Liddell's observations were of manual linguistic signs, not gestures, the two share the same universal kinesic affordances; the analogous phenomenon occurs in both because they share this "common ground" (see Kendon 2004, ch. 15). Gestural buoys are a particular kind of 'retrospective' hold: as with the examples discussed above, they usually originate as an ordinary gesture with a stroke synchronized with co-expressive speech, which remains in the gesture space as a post-stroke hold. But by remaining as a hold while new gestures are performed, gestural buoys can continue contextualizing new speech-gesture composites indefinitely. As McNeill (2005, pp. 178-179) has mentioned, speakers sometimes perform gestural buoys in a continuous, seamless alternation between the hands, a phenomenon he refers to as "layering with two hands." With each hold lasting until the hand is needed for a new gesture, every new speech-gesture composite is contextualized by an existing hold, in a fashion analogous to what a mountain climber must do while ascending a vertical face. While this mechanism is operating, every moment in the discourse exhibits embodied cohesion with its immediate past.[12]

[12] McNeill (2005, p. 178) also mentions an idea he attributes to Susan Duncan, which is that this layering phenomenon exhibits a division of labor between the hands in terms of which is carrying "discourse content" and which is carrying "depictive" or "object content." This characterization is meant to capture the notion that the longevity of the gesture which "enters hold mode" is what allows it to embody temporal cohesion across stretches of discourse. I would however caution that a gesture which is supporting discourse cohesion is not necessarily any less "depictive" because of it. Discourse cohesion arises from any hold of any length, coexisting peacefully with a gesture's object representational capacity.

In earlier writings, McNeill (1992, pp. 177-178) briefly mentions an example much like a 'buoy', in which an iconic shape is maintained continuously by the left hand through a series of descriptive statements that are each accompanied by a new right-handed gesture. McNeill notes that "the maintained handshape created a cohesive link uniting all the statements and other gestures," yet he also casts the phenomenon as a kind of "gesture repetition," which I believe downplays the fact that the hold is *retained* across multiple clauses. Referring to it as "repetition" implies that the speaker must repeatedly perform the continuation of the hold with each pulse of speech, and casts it as a sequence of identical holds lined up in series.

Example 2.9 — establishment of a set of landmarks by means of gestural buoys

The first row of images in this example establish the speech and gesture context leading into the speaker's use of a pair of gestural buoys, which are shown on the following page. The model village is absent from the interaction, and I have applied a few graphical aids to help clarify the spatial relationships and entities performed in gesture. Gesture holds giving rise to buoys frequently begin with symmetrical, two-handed gestures: one hand of a symmetrical hold is left in place while the other continues to perform new gestures, in a pattern Enfield (2004, 2009) dubs the "symmetry-dominance construction." This is apparent below, in the transition between image [c4] on this page and image [d1] on the following page. However, the second buoy in the example, which begins at the moment shown in image [d2], emerges from the post-stroke hold of a *one-handed* gesture. Thus, buoys can emerge in chains, and from single-handed gestures: the symmetry-dominance construction is a subcase.

(2.9) **B**: And House 33 and House 35, ·· **|A**: Mm-hm. ··

B: are ····· directly ········· where · towards the back?

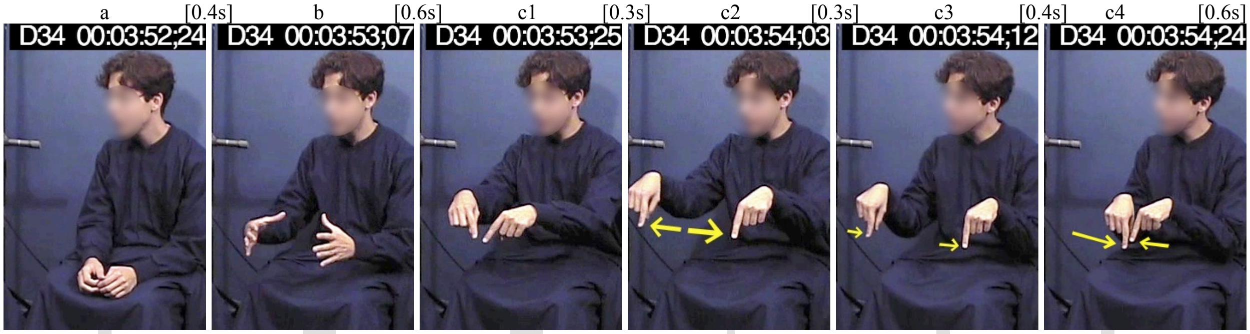
a b c1 c2 c3 c4
A: # Well there's like* · The whole town is like a square? ·· **|B**: Mm-hm.

d1 d2
A: and here's the church, and here's the train station, #

e1 e2 e3
here's the movie theater, · and here are the two houses. ·· **|B**: I see.

f1 f2 f3 f4 g
A: And there's just an intersecting street like that. ·· **|B**: Okay.

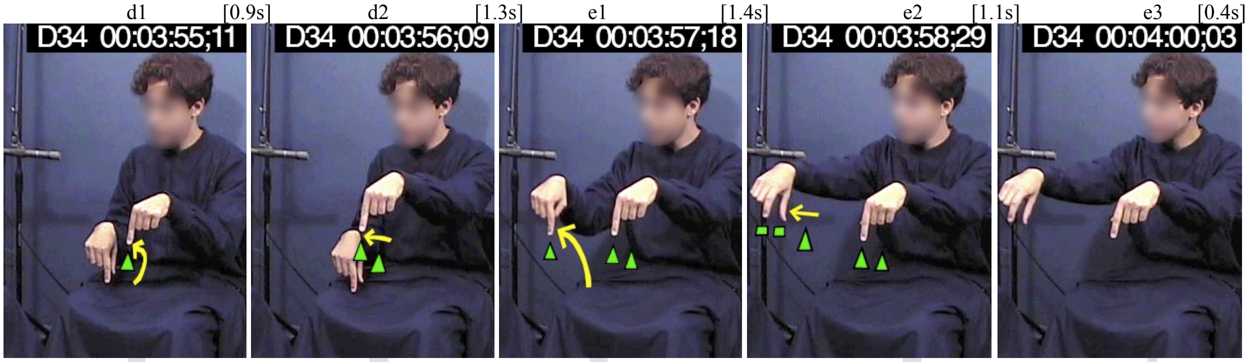
a-b: *BH rise from lap into loose symmetrical palms enclosing a basketball-sized space over his lap.*
b-c1: *BH then switch to palm-down fists with thumbs and index fingers extended straight down (shapes which are maintained throughout), as hands are brought together with arms extended outward.*
c1-c4: *Forearms move BH symmetrically away from each other, tracing a line tips of index fingers, then arms are pulled in to trace two lines perpendicular to the first, and finally the forearms bring the hands together again to trace a final line parallel to the first, after which the hands hold in place.*



A: # Well there's like* · The whole town is like a square? ·· **|B**: m-hm.
(BH) ↗ ↗

c4-d2: *While RH remains held in place for a total of 2 seconds (images* [c4] *through* [d2]*), LH lifts and moves forward a few inches to pinpoint and hold on a location in space, after which LH again lifts and moves forward to pinpoint and hold on another location in space.*

d2-e3: *While LH now holds in place for nearly 6 seconds (images* [d2] *through* [f4]*), RH repeats the pattern just finished by LH, first by breaking out of its 2 second hold to lift and move forward to pinpoint and hold on a location in space at the same height and distance from his chest as the current LH hold (image* [e1]*), then by separating index finger and thumb from each other as arm fully extends, to pinpoint and hold on a pair of adjacent locations still further away.*

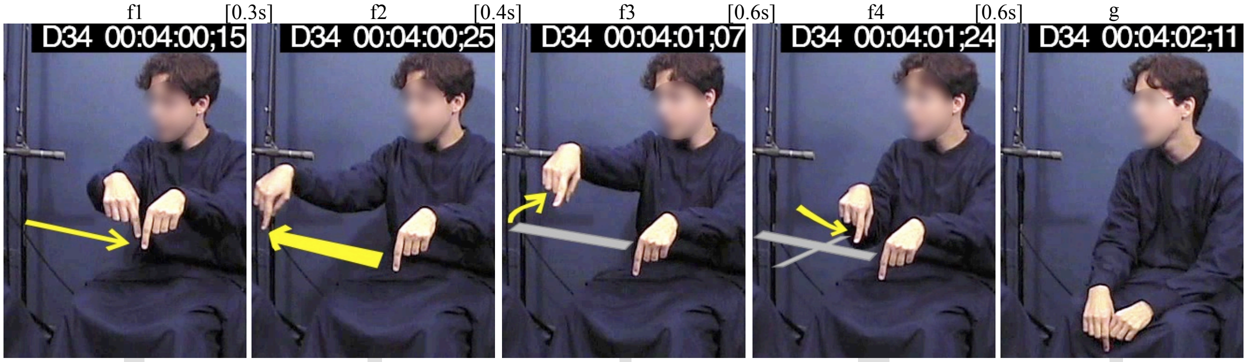


```
 A:  and here's the church, and here's the train station, # here's the movie theater, · and here are the two houses. ·· |B: I see.
(RH) ------------------------------------------  ↗ ʌ   -------------------  ↗   ʌ   ------------------------
(LH) ↗    ʌ ------------  ↗  ʌ    ----------------------------------------------------------------------------------
```

e3-f2: *While LH continues its long hold, RH brings thumb back against index finger to resume pinpointing with only the latter, as RH swings to his left to meet LH, then back again.*

f2-f4: *RH now moves out away from his body and slightly to his left to hold momentarily in the center of his left-to-right range, with arm extended to fullest extent. RH then draws in toward his body and past the still-holding LH, tracing a perpendicular line across the previously swept line.*

f4-g: *After maintaining this final configuration for a brief hold, BH drop to rest in lap.*



```
 A:     And       there's      just        an        intersecting          street     like     that.    ··  |B: Okay.
(RH)  ↗                        ʌ         ↗          ------   ʌ          -----------  ↘
                                                                                               ____________
(LH) ----------------------------------------------------------------------------------------  ↘
                                                                                               ____________
```

At the moment of image [c4] near the start of the example, the speaker has just completed a rectangular-shaped depiction of the borders of the village, performed with symmetrical movements of downward-pointing index fingers (images [c1] to [c4]). He then leaves his right hand anchored in place as a 'buoy' as his left hand pinpoints the virtual locations of two adjacent buildings (images [d1] and [d2]).[13] In so doing, his right hand maintains a remnant of the depiction of the town's borders, and this serves to spatially structure the building placements performed "inside" the borders by the left hand. His left hand then becomes the new buoy by strictly maintaining the position it achieved in image [d2], which is the position of the second and final building on that axis, the train station. It remains anchored there for nearly six seconds, all the way until the speaker drops both hands to rest at the end of the passage. During this interval, his right hand completes the placement of the buildings of the village, then pauses in place such that both hands remain anchored for 1 second (image [e3]) as he prepares to begin a new spatial description. This new description is about the streets of the village, not the buildings, but any currently anchored landmark is still congruent with spatial depiction of any other landmark, and so he does not lower his left hand until this second description is complete.

Example 2.10 — a continuous chain of buoys lasting across pauses and disfluencies

This final example illustrates several of the important points introduced at the start of the section. One is that with a *chain* of layered buoys, a speaker is able to keep a gesture hold in place at all times, without any disruption to utterance production beyond the constraint that new gestures must be performed with one hand at a time. By alternating roles, each new figural gesture can become the ground of the next, which can then become the ground for the previous limb's new figural gesture, and so on. In this way, each post-stroke hold can fulfill the role of 'ground' for as long a time as possible, before its limb is needed for a new gesture, and the pattern can be continued indefinitely.

Gesture buoys, like any retrospective, forward-contextualizing holds, can be maintained through disfluencies and pauses in speech. In this example the speaker has a certain amount of difficulty remembering the details of the task description he believes he must relay to his partner, and his speech contains numerous long pauses and asides. In support of these delaying tactics, and as a powerful defense against interference or the loss of his own train of thought, he maintains chained buoys through large swaths of his extended, hesitant description. These holds link together into the following gross structure: first, as his left hand remains buoyed in place, his right hand sweeps around in a very slow arc (images [b1] to [b3]), during which he begins smiling as he becomes amused and embarrassed by his failure to remember the details of the task. The slowness of his gesture arc both reflects and enables this delay. Then, leaving the right hand in place as the second buoy, he thrusts his left hand forward to represent the location of the trees behind the train station (image [c] on p. 63).

– (Main text continues on p. 64)

[13] In the images, the added small arrows indicate the path of movement of the gesturing hand(s) from one image to the next during most of the passage. Additionally, the small shapes show the virtual placement of each building, which I have propagated forward into subsequent snapshots to show the relative positions.

(2.10) **A**: The train ·· # pulls up ··· past a church and stops at the station. ··

B: Okay.

A: # ···· I · [a1]get · off ·· the train ·· at the statio[a2]n, go ar[b1]ound the station [b2]···· # [b3]····
and · there are two [c]trees. ··
I go bet[d1]ween the [d2]two trees, # and · there [d3]are · two hous[e]es. ··

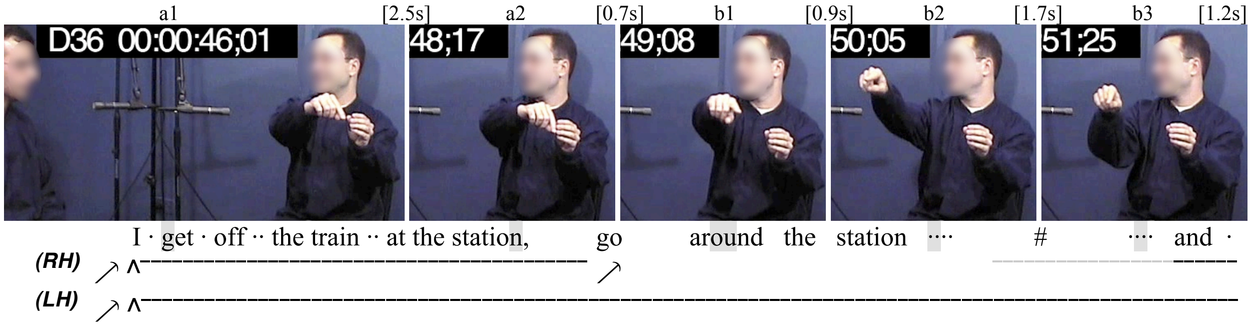
B: m-hm. ··

A: House thirty-[f]three, ··· and hou[g1]se thirty-five. ··

B: Okay. ··

A: Alright. ·· # I · believe · I··· {*swallow*} uhh ··
[g2]pick [h]up · assistants · there ··· # so I'm supposed to [i]ask the people in 3[j1]3 and 35
··· # how ·· to uhh* · or, [j2]t*f[k]or · assistant[l1]s ····
so · presumabl[l2]y we h[m]ave more than just u[n1]s · # uhh ·· that can do this ··
# and · # [n2]we then· proceed [o1]across the [o2]street ··· to·· the·· movie theater ··
the abandoned movie theate[o3]r [p]·· important piece of information. #

a1: *Leading up to this image, BH have swept forward from L shoulder along a path representing the arrival of the train, followed by a rise-fall stroke during which RH switches to an index finger deictic pinpointing the (figural) person disembarking from the LH (ground) of the train.*
a1-a2: *This configuration is held in place for 2.5 seconds of speech.*
b1-b3: *LH continues to hold in place as a buoy to RH's slow horizontal arc which extends away from his body and then back toward his R shoulder. During the last phase of this arc, RH's movement slows nearly to a stop as he pauses his speech for two seconds, but it very slowly continues toward his shoulder as he takes a breath and smiles at his interlocutor* (b3), *indicating embarrassment at how long he is taking to remember the next step in the path description.*

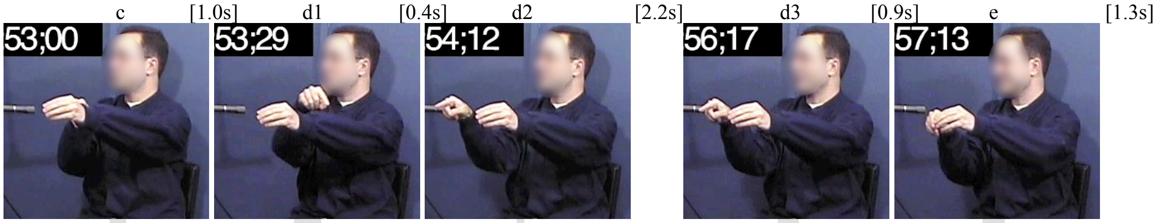


b3-c: *RH now becomes a buoy and remains in place as LH ends its own buoy-hold (which has lasted for 6 seconds) and thrusts forward and to his right to indicate the position of the trees.*

c-d2: *LH then holds in place as a buoy again, as RH resumes its figural role in pinpointing the path of the speaker's movement past the trees, via a sweeping stroke outward and to his right.*

d2-d3: *BH hold in this configuration for more than two seconds as he decides what to say next.*

d3-e: *RH then becomes a curled closed 5, matching LH, and BH enact a small downward stroke and hold in place through the interlocutor's backchannel contribution.*

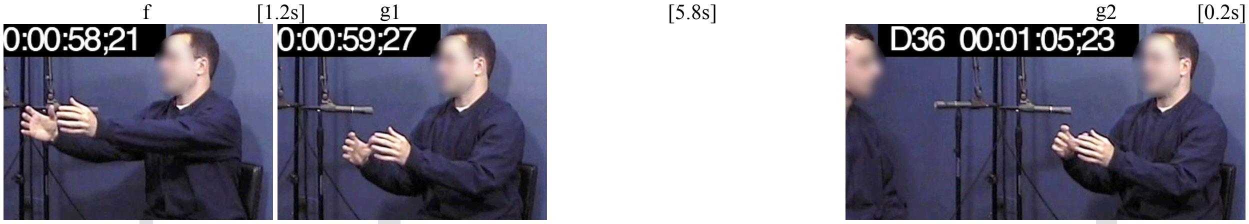


there are two trees. ·· I go between the two trees, # and · there are · two houses. ·· | **B**: m-hm. ··
——————— ↗ ʌ ——————————— ↗ ʌ ———————
↗ ʌ ———————————————————— ↗ ʌ ———————

f: *BH extend outward to full arm's length in a downward stroke (loose 5s, palms facing, 6 inches apart) to show the location and approximate size (in the model) of the most distant house.*

f-g1: *After a brief hold, this is repeated a foot closer to his body, which places the neighboring house.*

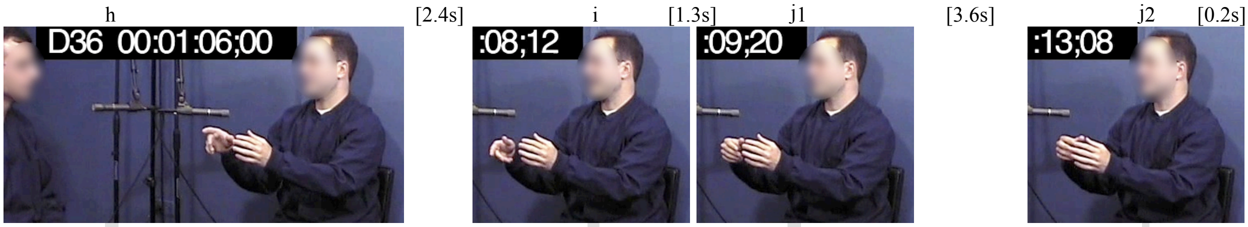
g1-g2: *He then pauses in speech and remains frozen in place for nearly 6 seconds, through his partner's backchannel acknowledgment and his own time-filling but largely vacuous speech and further pausing. Fluent speech and gesture return together on the word "pick" (image* [g2]*).*



A: House 33, ··· and house 35. ·· |**B**: Okay. ·· |**A**: Alright. ·· # I · believe · I··· {*swallow*} uhh ·· pick
↗ ʌ—— ↗ ʌ ———————————————————— ↗
↗ ʌ—— ↗ ʌ —————————————————————

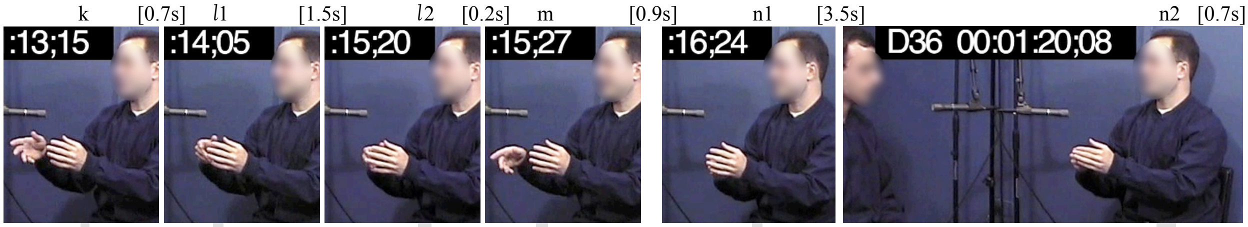
h-j1: *LH continues to hold in place as a buoy while RH enacts small, abstract deictics that modulate the steps in the subtask of recruiting assistants, each of which holds in place until the next.*

j1-j2: *BH hold in place symmetrically for 3.6 seconds, extended by a period of disfluent speech.*



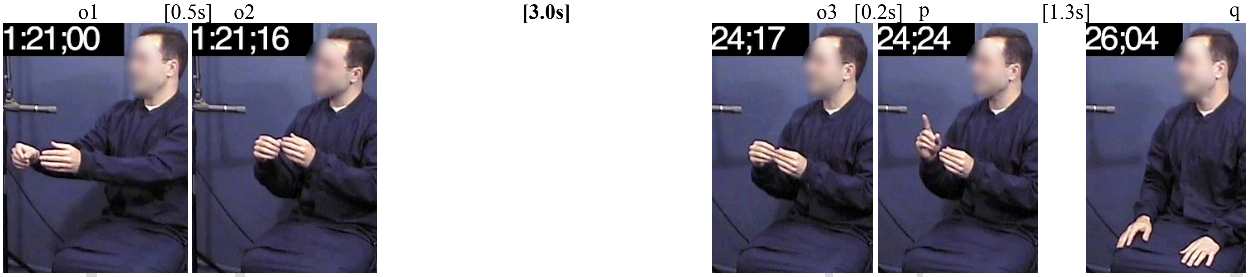
up · assistants · there ··· # so I'm supposed to ask the people in 33 and 35 ··· # how ·· to uhh* · or, t*
ʌ ——————————— ↗ ʌ ———— ʌ ————————————
———————————————————————————————————————

k: *Fluent speech returns along with another palm-up excursion (similar to image* [i]*).*
l1-l2: *BH return to a symmetrical hold across a 1 second pause.*
m-n1: *RH points toward his partner and then back toward himself, resuming the BH symmetry.*
n1-n2: *The now-resumed BH symmetrical hold is maintained for 3.5 seconds.*



for · assistants ···· so · presumably we have more than just us · # uhh ·· that can do this ·· # and · # we then

n2-o2: *LH has been buoyed in place for **over 20 seconds** (images* [g1] *to* [n2]*) but now finally breaks the hold as BH extend in preparation for an inward pulling stroke representing the motion of the assembled party from the houses to the theater.*
o2-q: *Back in the familiar BH symmetrical hold, but at this new virtual location, the hold is kept for 3 seconds before RH index finger extends upward to emphasize the "important piece of information." As he finishes this final sentence he finally drops BH to rest on thighs.*



proceed across the street ··· to·· the·· movie theater ·· the abandoned movie theater ·· important piece of information #

(Continued from p. 61) – While holding his left hand in place as the second buoy in the chain, he sweeps his right hand forward to represent, from an observer viewpoint, his own movement between the trees (images [d1] and [d2] on p. 63). At this ensuing hold, his gestures return to being symmetrical and both-handed, through the placement of the virtual location of House 35 (image [g1]). This marks the beginning of a massive left-handed buoy lasting for over twenty seconds, as his right hand performs small, non-representational (metanarrative) gestures and he goes through several disfluent asides regarding the tasks to be performed at that location. The fact that these gestures are non-representational means they do not depend on the buoy for their contribution to the interaction, unlike in the typical symmetry-dominance cases discussed by Enfield (2004, 2009; see also the first buoy in Example 2.9 above). However, the buoy still maintains temporal continuity across an interval of speech and gesture forming a unified subtopic.

Even if the precise spatial location of the gesture loses its relevance over time, the lack of change of position matches the lack of shift of topic.

I question whether the maintenance of the hold requires “a persistent expenditure of effort,” as Enfield believes (2009, p. 124), or even carefully monitored “full control” (Kendon 2004, p. 137). It could instead be the lack of shift to a new topic which allows the body to keep a consistent configuration, and the position of the limb might not be carefully controlled until such a shift occurs (other than whatever background effort is required to maintain muscle tension against gravity). Just as we would assume that the maintenance of a consistent posture across the prosodic “tone groups” of a complete “locution” (see Kendon 1972) would require little or no conscious effort, compared to an active shift in posture, I believe we should seriously investigate whether maintenance of the limbs can be achieved with similarly little in the way of actively directed control.

At any rate, by the time we reach image [n2] (p. 64), the position of the extended left hand buoy no longer represents the location of House 35. This is made clear by the fact that the speaker reestablishes the location of the house(s) by reaching out to a new location with both hands, representing them as a starting point (image [o1]), and then draws his hands back toward himself to show the movement of his team crossing the street on their way to the movie theater. After a final left hand buoy during another right-handed metanarrative gesture, he finishes his task description and drops both hands to his lap.

In this example, the ‘chain of buoys’ pattern seems to treat the gesturing hands analogously to the placement of external objects in a scene: each item can be manipulated and left in a position until needed again at a later time. The limbs of the body could be subject to a sort of ‘gumby effect’, in which a decision to *dis*engage their position might require more effort than simply keeping them in place. McCullough (2005, p. 5, footnote 7) appears to support a similar idea: with gesture holds, referential maintenance is simply accomplished, with “no additional effort” required beyond whatever resources are needed to maintain the gesture’s shape and position in space, against gravity and the body’s natural inclination to return to a relaxed state. My motivation in stressing this possibility is that I wish to emphasize that gesture holds, as temporary artifacts of position and/or shape information, could play the same sort of role in ‘distributed cognition’ as the external artifacts and structures emphasized by Hutchins (1995, 2006) and Goodwin (2000, 2006). Just because body-based artifacts are attached to the same flesh that houses our thinking minds, does not mean they can receive any special exemption. Instead, they are the most convenient and malleable of artifacts, always available to assist in cognitive and communicative tasks. I return to this theme in Chapter 4.

Conclusion

In this chapter I have presented a survey of some of the evident prospective and retrospective functions of gesture holds coinciding with disfluencies and pauses in speech. Gaps and breakdowns in the flow of the spoken part of an utterance are frequently accompanied by a kind of slowing or stasis of gestural movements which, especially from the listener's point of view, continue to support 'delivery' of some of the speaker's intended message. A gesture hold embodies a temporal span during which the elements of the multimodal composite, produced together but sometimes seemingly unbound because of a hiatus in speech delivery, remain associated across time. This association can bind present to future, as in the case of a gesture whose expression creates a 'prospective' relationship with speech, or past to present, as in the case of a gesture that maintains a retrospective link to an earlier moment. In combination, the present is bound to both past and future, as when a gesture spanning a hiatus is retrospective while also serving as the basis for future speech. With each utterance considered as a multimodal composite, the *edges* of the utterance 'move' appear to become frayed, rather than unified and tidy, with content expressed in an overlapping manner that creates local cohesion of discourse, binding the current state of an interaction with both its future and its past.

In the next chapter, I continue to discuss the prospective effects of gesture holds occurring at the start of disfluent utterances, with a specific focus on how my findings interface with (1) other work on gesture and disfluency, (2) work suggesting that certain disfluencies fulfill important speaker goals which can be supported much more powerfully when gesture is taken into account, and (3) work supporting my claims of the rapid uptake and referential power of the gestural cues I have been illustrating.

3 Gesture holds and listener uptake during preliminary utterance commitments

In Chapter 2 I presented a survey of gesture holds occurring during disfluencies and pauses in speech, which I claimed could support prospective as well as retrospective referential functions from the point of view of a listener attempting to make sense of a speaker's message. The chapter did not, however, discuss any of the existing literature on gesture and disfluency and how it compares to my own examples, a topic I now briefly explore.

Next, I discuss research outside of gesture studies which has proposed that speech disfluencies can serve important functions for speakers as well as listeners. For example, speakers often pause just after the start of utterances, evidently allowing them to begin speaking before the utterance is fully formulated. Listeners, meanwhile, appear to make use of such patterns by orienting toward new potential referents rather than recently mentioned ones. As I will show, these findings become even more compelling when simultaneous gestural cues are taken into account.

This leads me to a discussion of the pitfalls of the perennial debate on speech-gesture synchrony, or lack thereof. The debate has hinged on the belief that synchrony between the modalities should be judged entirely on the basis of the timing of whatever lexical item from surrounding speech seems most redundant with gesture. However, speech which is co-timed with the stroke of a gesture is often synchronously co-expressive even when the content expressed in the two modalities is broadly complementary, a fact which is independent of whether the gestural content is also reiterated in later speech. Still, it is worth exploring the question of whether a fully synchronous, co-expressive pulse of an utterance may be internally asynchronous in terms of a listener's uptake of content from the respective modalities. Toward this end, I illustrate an example of the compounding effects of prospective gestural cues that are also recurring discourse 'catchments' (see McNeill 2005, ch. 5). Due to the recurrence of particular linguistic items with a particular iconic form, the example illustrates a plausible mechanism whereby a listener could be led, by gesture, to accurately predict impending linguistic items during the earliest moments of an utterance, at the start of the gesture's preparation phrase.

Previous work on gesture and speech disfluency

With each of the works discussed below, I aim to briefly mention what the work was responding to in its academic context and what it added to our knowledge on the subject of gesture and speech disfluency. I also discuss any evident problems or complications suggested by comparison with my own examples presented thus far.

McNeill 1992 — gesture rates during fluent versus disfluent speech

One of McNeill's goals in his 1992 book was to emphasize that the great majority of gestures occur as part of normal fluent spoken utterances, rather than during silence, in order to counter claims that the main impetus behind the production of gestures is that they are triggered by breakdowns in speech. Butterworth and Beattie (1978) had found nearly identical rates of gestures per 1,000 seconds of speech and 1,000 seconds of hesitation, but McNeill (1992, p. 92) gives a table showing that 90% of gestures occur during co-articulated speech, with only 10% occurring during disfluencies of various sorts: 1% during filled pauses, 2% during unfilled pauses, 3% during breath pauses (usually inhalation), and 4% during "false starts." McNeill succeeds in making the point that, even if gestures occur at roughly equal rates during fluent speech versus speech hiatuses, the former makes up the vast majority of the time spent in narrations—and therefore the vast majority of gestures—and so it makes little sense to claim that speech breakdowns are the primary impetus driving gesture production.

However, McNeill's figures don't tell us anything about what the remaining 10% of gestures are accomplishing during speech breakdowns. Furthermore, his statistics used the gesture *stroke* as their only criterion for calculating what a gesture is simultaneous with. As I showed in Chapter 2, gestures whose strokes occur with fluent speech can nonetheless fulfill important functional roles during disfluencies, pauses, and self-interruptions, when they include a 'retrospective' post-stroke hold spanning across the more problematic intervals. Meanwhile, Example 1.1 from Chapter 1 could possibly be interpretable as a gesture whose *pre*-stroke hold spans a disfluency (and may lack a stroke entirely), with only the next gesture's stroke coinciding with fluent speech. This means that the percentage of gestures which may be functionally significant during disfluent speech could be well above 10%, if we add these cases to those making up McNeill's figures. Other examples presented in Chapter 2—those 'prospective' cases whose strokes do not coincide with fluent speech—would presumably belong to his 10% of gestures whose strokes coincide with false starts, pauses, and so on. In all cases, I have emphasized the potential ability of these gestures to continue revealing the speaker's intended message to a listener, in spite of the speech hiatus. My emphasis is mostly due to the fact that I, as an analyst reviewing the taped conversations, am closest to the role of a third party listener. Similarly, the reader, when interpreting transcripts and sequential snapshots, is closest to matching the listener's role in the original conversation and can somewhat gauge the plausibility of arguments oriented to that perspective. Of course, it may be the case that gestures occurring during disfluencies do help a speaker resume fluent speech, and that they also affect the linguistic choices in the resumed speech, but their simultaneous communicative utility suggests that they are not simply *caused* by disfluency, especially given that they seem otherwise very similar to gestures performed during fluent speech.

Kita 1993 — persistence of gesture during speech repetitions

In the 6th chapter of his 1993 dissertation, Kita addressed the question of whether speech and gesture unfold via interactive, mutually accessible processes after they are deployed, or whether they unfold within "impenetrable" modules. To study this question he analyzed gesture behavior during 135 cases of self-interrupted speech (from narrations by native English speakers) and compared it to a baseline established from fluent utterances in the same corpus. Although these self-interruptions form just a subset of what I have broadly called disfluency or hiatus, Kita's findings turn out to be very relevant to my discussion.

He classified the self-interruptions in his data as either *repetitions* or *repairs.* In repetitions, the last word or words prior to an interruption are repeated, without change, to start the resumption. Levelt (1983) referred to these as "covert repairs," but Kita found evidence to keep them entirely separate from Levelt's other repair types.[1] What Kita found was that among self-interruptions in speech that coincided with a gesture phrase, the gesture phrase terminated before the beginning of the resumption in 61% of *repairs,* but terminated before the resumption in only 16% of *repetitions.* Both findings were significantly different from a baseline derived from fluent speech, in which the figure stood at 35%. In other words, during repairs, speakers have a tendency to drop their gesture phrase to rest or start a new one, at the resumption. But in repetitions, speakers have a tendency *not* to abort or shift to a new gesture, and instead continue the gesture phrase already in play at the time of the repetition.[2]

Many of the examples I have presented thus far do not fit Kita's self-interruption criteria, because they involve speech pauses that simply continue after the hiatus without repetition or repair. But see Example 1.1 in Chapter 1, in which the first part of the disfluency involves repeated words ("we'll we'll…") along with the continuation of a

[1] These other types include "fresh starts," in which the resumption following the interruption does not continue the earlier utterance, "retracings," in which some of the earlier material is repeated with one or more modifications to begin the resumption, and "instant replacements," in which the last word before the interruption is replaced with something else that then continues the utterance.

[2] Seyfeddinipur (2006, p. 94) notes that these figures should be taken with some caution, because one of the factors likely to influence a gesture's tendency to have been terminated is the amount of time elapsed between the gesture's onset and the point of measurement. Since self-interruptions tend to involve a suspension of the utterance, repairs may have a greater tendency for gesture termination than the baseline simply because they generally involve a time-spanning suspension prior to the resumption. However, this observation should not affect Kita's finding that gestures tend toward *continuity* (compared to the baseline likelihood) in the case of repetitions, because it seems exceedingly unlikely that repetitions would on average involve a significantly shorter interval before resumption, compared to the baseline speaking rate. In fact, Clark and Wasow (1998, p. 221) show that repetitions are far more likely after a filled or unfilled hiatus than they are during undisrupted delivery, which further validates Kita's figures.

gesture hold, followed by a definite 'fresh start' in which the gesture switches to a two-handed thrusting stroke that is arguably a new gesture phrase. See also Example 2.2 in Chapter 2, in which the speaker repeats nearly the same word continuously ("they're they're they're they're… they… they…"), and maintains a single gesture phrase throughout. Also fitting Kita's self-interruption criteria are examples presented later in the current chapter, in which a gesture hold persists through a single repetition (see especially Examples 3.1 and 3.2).

Kita found a major difference in the linguistic material occurring in repairs versus repetitions: 71% of repetitions occurred with "discourse-determined" words such as articles, pronouns, and conjunctions or other connectors, while 29% occurred with "content-determined" words such as nouns, verbs, adjectives, and prepositions. With repairs, meanwhile, 32% occurred with "discourse-determined" words, and 68% with "content-determined" words. These figures line up quite well with Clark and Wasow's (1998) findings on the types of words most frequently associated with preliminary utterance commitments that undergo repetition. But even more crucial are Kita's observations on the timing of gesture with speech repetitions: he found that repetitions coincide significantly more often with the *preliminary phases* of a gesture phrase (with "preliminary" here defined by Kita as either preparations or pre-stroke holds), compared to both repairs and the fluent baseline. Therefore, when I argue that there is listener uptake of gestural content during speech repetitions, I am often arguing for listener uptake from these same preliminary gesture phases, prior to the stroke. As I will illustrate shortly, the inclusion of such gestural cues can greatly enrich Clark and Wasow's (1998) model.[3]

Mayberry and Jaques 2000 — gesture during stuttered speech

This work aimed to investigate the degree of integration of gesture and speech by examining gesture behavior during pathological stuttering. All of the examples I have presented involve "normal" disfluencies, as opposed to stuttered disfluencies; Mayberry and Jaques (2000) distinguish the latter by the presence of repetitions and prolongations at the level of individual syllables and sounds, an effect which is absent from normal disfluencies (a possible instance of true stuttering occurs in Example 3.5a below). While gesture behavior during stuttered disfluencies appears to be rather different from gestures occurring during normal speech (whether fluent or disfluent), there are some similarities.

[3] Consistent with my stance that gestures during repetitions are already 'delivering' valid utterance content, Kita disagrees with Levelt's (1983) belief that repetitions are simply a form of repair in which an oncoming speech error is detected and fixed before actually appearing in speech. Instead of involving an actual *error* committed and in need of correction, Kita believes repetitions involve an entirely different psychological process: one of "miscoordination" between discourse-level thinking at the very start of an utterance, and "imagistic" or content-level thinking that must be transitioned to. The discourse-level thinking required for each utterance is a direct consequence of the need to string sequential utterances together into a coherent discourse, and a hiatus with repetition can occur "when discourse thought finishes encoding too early, or imagistic thought fails to encode in time" (Kita 1993, p. 79). This is basically consistent with Clark and Wasow's (1998) theory, which holds that speakers sometimes make preliminary commitments in order to fulfill an expectation to speak and continue the flow of discourse. However, Clark and Wasow would not characterize this as "miscoordination."

The authors found that gesture strokes virtually never occurred during intervals of stuttering, but did occur with the return of fluency. Although stuttered utterances were in general accompanied by very few gestures, those with gestures exhibited the following behavior, in order from most frequent to least frequent: (i) the hand would fall to rest at the onset of stuttering, but return to the gesture space immediately upon resumption of fluent speech; (ii) the hand would stop moving and *hold in place* at the onset of stuttering, and resume movement immediately upon resumption of fluent speech; (iii) the hand would fall to rest at the onset of stuttering and stay there even after the return of fluent speech. The second of these, in which the hand holds in place during stuttering, seems closely related to my suggestion that normal disfluencies may involve 'embodied stasis as a mechanical response' (see Example 2.1 on pp. 37-39).

Mayberry and Jaques concluded that the effect of stuttering on gesture offers support for claims of the close integration of gesture and speech. For my purposes, their research suggests that gesture holds are not usually present as a possible mitigating factor spanning stuttered disfluencies. In the very few cases where a gesture is not dropped during stuttering, however, a gesture hold *is* forced, because the gesture stroke cannot occur until the return of fluent speech. In their example (Mayberry and Jaques 2000, p. 207), a speaker raises his hand into a flat palm facing down, then freezes into a stuttering-induced, extended pre-stroke hold that spans many repetitions of the initial phoneme /r/. Eventually, the fluent utterance becomes "ran across the road," accompanied by the hand sweeping back and forth in a stroke. Mayberry and Jaques intended this example to illustrate that representational gestures never occurred during stuttering, but they appear to miss the representational possibilities of the non-stroke portion of the gesture—namely, the flat handshape given by the extended pre-stroke hold spanning the entire stuttered interval. Such a gesture hold may or may not have an effect on the speaker's attempts at retaining coherence through the stuttered interval (see Chapter 4), and I have also argued that in context, such disfluency-spanning holds can potentially help the listener orient toward particular expectations regarding the likely semantic scope of the impending fluent utterance. If this is the case, it is a mechanism that can operate during stuttered as well as normal disfluencies.

Seyfeddinipur 2006 — self-interruption of gesture and speech

Following on Kita (1993), this work directly examined the relationship between gesture suspensions and speech self-interruptions, in order to discover whether cues visible in gesture could be an indicator of the moment of error detection. Since gestures tend to slightly precede semantically co-expressive speech, changes in gesture could be detectable signals of "covert" error detection ahead of actual speech interruption and repair. Seyfeddinipur found that gesture suspension often accompanies speech disfluency and tends to slightly precede it, by roughly 100 milliseconds. However, gesture suspension is also a frequent accompaniment of fluent utterances, and she did not find any significant difference in the rate of gesture suspension in fluent versus disfluent utterances.

While this work had an object of study that was arguably very similar to the theme of the current chapter and Chapter 2, I faced a number of difficulties when attempting to integrate Seyfeddinipur's findings with my own investigations. One minor difference is that I am interested in the functions of gesture holds during any kind of hiatus in speech, including silent pauses in delivery that are arguably not "disfluent." Seyfeddinipur includes cases of repair, repetition, or insertion of a non-silent hiatus (such as "uhh…"), but does not include other instances of hiatus. A second difference, much more problematic for making comparisons with my findings, is that while a gesture hold certainly involves a configuration being "suspended" in space and time, this is not what Seyfeddinipur means by gesture "suspension." In her investigation, a suspension is a discrete transition ending one phase and beginning another—that is, a gestural self-interruption—which is quite distinct from my focus on gesture holds as time-spanning intervals during which a configuration is kept in stasis. In Seyfeddinipur's data, the moments at which hold phases begin are just a subset of the aggregated count of gesture suspensions, which can consist of (i) any transition from a dynamic (movement) phase into a static phase (i.e., a transition into a hold), (ii) any transition into a retraction of any kind, (iii) any case of a preparation being followed by another preparation, or (iv) any case of a preparation or stroke being interrupted mid-stream and followed by something else. These suspensions are the transitions referred as "stop shifts" by Seyfeddinipur and Kita (2005). Unfortunately, while there are tables giving the number of suspensions marking the end of each type of gesture phase in disfluent utterances versus in a simulated fluent baseline, no figures are provided showing the rate of specifically those suspensions characterized by transitions *into gesture holds,* in disfluent versus fluent utterances. Thus, while an answer might be forthcoming if we were to conduct a new analysis of the data, Seyfeddinipur (2006) does not yet address the question of whether speech disfluencies affect the likelihood of gesture holds.

Furthermore, even supposing that the frequency of holds is not greater in disfluent utterances, it could still be the case that disfluencies cause holds that *last longer* than they do in fluent utterances. Seyfeddinipur (2006, p. 98) notes that many gesture suspensions preceding disfluent utterances are the same type as those often preceding fluent utterances (such as pre-stroke holds), and that we therefore cannot assume that such suspensions are caused by the disfluency. I have proposed that exactly such holds—those which would have been equally appropriate in a fluent version of the utterance, could prove to be lengthened in time due to the insertion of a hiatus in the flow of speech, and that this longer lifespan may have functional significance. During the hiatus, a prolonged hold may provide referentially significant information to the listener (and could also serve important functions for the speaker, as I will discuss in Chapter 4). The functional utility for the listener seems especially plausible given that disfluencies have a tendency to draw the gaze of listeners to the speaker (Goodwin 2006). Unlike Seyfeddinipur, I am less concerned with the question of whether "error detection" in speech triggers a "stop shift" in the current gesture phase, than I am with the retrospective and prospective referential potential of hiatus-spanning configurations when they do occur. It would be very interesting to find out whether time-spanning configurations are, on average, more frequent or lengthened in time when occurring with speech hiatuses, compared to fluent speech. As mentioned above, we have some suggestive evidence from Kita's (1993) finding that gesture phrases tend not to terminate during the hiatus associated with repetitions (the specific disfluency referred to as “covert repair” in Seyfeddinipur’s work, following Levelt 1983). A more comprehensive answer will have to wait for future research.

Clark and Wasow’s commit-and-restore model

When speakers repeat a word, they are not necessarily doing so because of an “error” in their speech planning, nor even because of a “miscoordination” between speech designed for discourse cohesion and speech designed for content, as Kita (1993) suggested. Clark and Wasow (1998) propose that word repetitions occur instead as a natural consequence of competing conversational pressures. One of these pressures may be to begin speaking as soon as possible in order to retain control of the floor and maintain continuity with previous utterances. Another pressure, however, may be to speak using syntactic constituents that are continuous (that is, hiatus-free). A hiatus might therefore be inserted right after the first word of many constituents, with that first word serving as a *preliminary commitment*. The word is then repeated to begin the resumption once the rest of the utterance is formulated, creating an undisrupted final version of the constituent. Just as Kita (1993) reported, Clark and Wasow (1998, p. 211) found these repetitions to be much more common with function words such as conjunctions (30.8 repeats per thousand), pronouns (37.7 repeats per thousand) and determiners (28.8 repeats per thousand) than with ‘content words’ (2.4 repeats per thousand).[4] Crucially, function words tend to be the “left-most” (that is, earliest) words of constituents such as noun phrases, verb phrases, prepositional phrases, and clauses, whereas content words tend to occur later in the constituents.

The authors present three lines of evidence for their model. First, in support of a ‘complexity hypothesis’, they show that speakers are much more likely to repeat an initial “the” or “a” of a noun phrase when it is complex than when it is simple, and they are also more likely to do so when the NP is at the left edge of a large constituent (e.g., a clause) than at the left edge of a small constituent (e.g., the object of a verb or preposition). Second, in support of a ‘continuity hypothesis’, speakers are much more likely to insert a pause just before beginning a constituent than after its first word. Simultaneously, when a word is repeated to restart a constituent, a delay is far more likely to occur right before the resumption than right after it has begun. Speakers are also more likely to repeat a word, the more lengthy and disruptive the hiatus, presumably because a more disruptive hiatus inserted into a constituent makes it more difficult to treat that constituent as continuous. The third line of argument is the ‘commitment hypothesis’, by which the authors claim that speakers make preliminary commitments to utterances even when they are not ready to produce all of the necessary linguistic material. In support of this they show that many repeated words have their first instance pronounced as a ‘phonological orphan’, such as “the” pronounced as “thiy” (i.e., “thee”), and that such words also tend

[4] Clark and Wasow (1998, p. 209) define ‘content words’ as those lexical items referring to “entities, events, states, relations, and properties in the world. They are characteristically nouns, verbs, adjectives, or adverbs.”

to be ‘near repeats’ rather than exact repeats, such as when “a” is used before the hiatus but then switches to “an” when required in the resumption. In other words, speakers begin some constituents while already anticipating that they will immediately suspend speech.

It turns out, of course, that speakers are often gesturing when they make these preliminary commitments, often in the form of a hold which spans the hiatus and is also part of the resumed utterance, just as Kita (1993) found. As discussed previously, these gestures are not only signaling an intent to continue, they are also already revealing information about what the continuation will be about. One of Clark and Wasow’s (1998) claims is that speakers must be able to estimate, at some conceptual level, the complexity of the impending constituent that they suspend. We can combine this with McNeill’s (1992, 2005) claims regarding the co-emergence of speech and gesture, in which a holistic image is part of the Vygotskyan (1986) ‘psychological predicate’ (the relevant departure from context which spurs the utterance): it could be the case that the activation of the image during the first stages of utterance formulation is crucial to a speaker’s ability to estimate the utterance’s complexity.[5] This image, emergent with the first word of the constituent just before the hiatus, can then become partly accessible in gestural form to addressees. In such cases, the pre-hiatus utterance is not just a preliminary commitment, it is also a preliminary *fulfillment* of that commitment, which strengthens Clark and Wasow's position even further. They note that, “as Jefferson (1989) found, speakers won’t tolerate a pause mid-utterance that is more than about one second long. When speakers anticipate too long a pause, they need to deal with it. They can use a filler, editing expression, or preliminary commitment to the next constituent” (Clark and Wasow 1998, p. 238). But as I would obviously like to emphasize, they are also often *using gesture:* in the speech stream, they can put forth a preliminary commitment in the form of a function word, meanwhile a co-timed gesture is already providing some of the content.

Example 3.1 — pre-stroke holds create early fulfillment of preliminary commitments

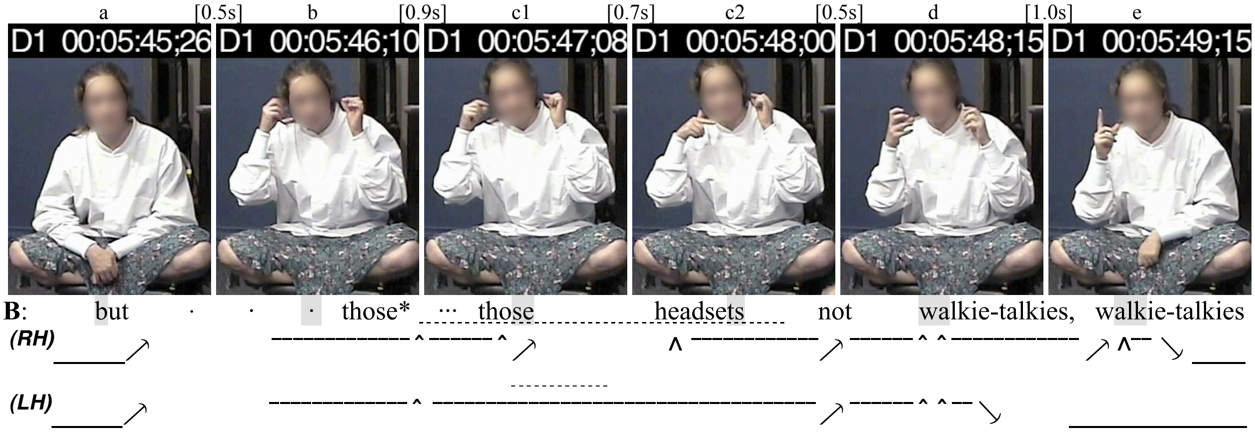
The example below illustrates this combination clearly. Participant B repeats the first word, “those,” of what will become a fluent utterance pulse consisting of “those headsets not walkie-talkies.” But just before the first instance of “those” is spoken, a pre-stroke hold is already in place that gives some basic information about the referent expected as the object (see image [b]). A linguistic cue to this referent must eventually finish emerging in linear-segmented speech, but the referent already has a referential presence, via gesture, that goes beyond the preliminary commitment from speech. The demonstrative determiner (“those”) and the gesture, taken together, effectively serve as the full speech constituent’s proxy while also contextualizing it ahead of its full delivery. As I discuss following the example, the gesture hold spans across (and is therefore co-timed with) several elements of the speech stream, and synchronizes logically with all of them.

[5] The term “image” here is not meant as a simple picture, but rather a transformation of a scene involving a variety of spatial, motoric, and other parameters.

(3.1) **A**: And once you hear the rowdy Australians, then ·····
or sh* do you think we should have walkie-talkies maybe ···
[**A**: Eh in case you can't he*]
[**B**: # Well we definitely need walkie-talkies ··]

a b c1 c2 d e
[**A**: Aah! Very good. Ok*]
[**B**: but ··· those*··· those headsets not walkie-talkies, walkie-talkies are too noisy.]

a-b: *BH rise together into symmetrical pinching shapes at the ears, which are held in place.*
b-c1: *Just after the first "those," BH begin waggling emphatically through the silent hiatus.*
c1-c2: *LH stops waggling and buoys in place as RH converts to an abstract index finger deictic, with a main stroke on the first syllable of "headsets" and repeated small waggle-thrusts after the stroke.*
c2-d: *BH change together into symmetrical grasping shapes, with a tiny double clench on "walkie."*
d-e: *LH drops to rest, but RH performs another abstract index finger deictic, this time pointing straight up and timed with the second instance of "walkie"; RH then drops to rest as well.*



During the portion of the utterance spanned roughly by images [b] and [c1], the gesture is representative of the syntactic object of "those," yet by virtue of being in a separate modality it can be performed simultaneously with the determiner. During the hiatus, the gesture continues to represent the referent the speaker has committed to, which she will therefore continue speaking about. Then, around the moment of image [c2] the gesture (or half of it anyway) continues to represent the noun which is itself now spoken overtly in speech. In all cases, the modalities are associated through synchrony: a determiner with aspects of its object, a gap in speech with an embodied commitment to continuity, and a noun with (once again) some of its aspects represented visuospatially. By recognizing her immediate goal as being the task of describing a referent, these elements are all also associated with each other across a slightly longer timespan.[6] We might separately

[6] Sidnell (2006, p. 404) believes the recognition of particular goals and activities plays a role in overcoming the "binding problem" that Levinson (2006, p. 53) believes exists regarding our ability to differentiate which elements of multimodal signals are meant to be associated, similar to Levinson's own assertion that some kind of "goal analysis" might be needed.

ask why the information from the two modalities is combined in just this fashion—for example, how is her gesture understood as coreferential with the expected syntactic object of "those," rather than combining with speech via some other association? But this question is much larger than my investigation: the same question already exists for many sequential signs within a single modality, whose associations can be left relatively unconstrained by grammar (consider, for example, the flexible associations between compounded nouns in English, or between nouns and noun-modifying constructions in Japanese: see Downing 1977; Matsumoto 1997).

Regarding the listener's uptake in the example, during the hiatus he should already be quite sure of the following: (1) the referent is a kind of walkie-talkie or has a similar function, because the speaker has just said "well we definitely need walkie-talkies but those…," (2) the referent should be familiar to him from experience, due in part to the speaker's use of "those" instead of other choices, and (3) one of the referent's major qualifying differences from his earlier suggestion is that it is a device placed at the ears (this is information found only in gesture). This last piece of content begins to be perceivable a half second before the first instance of "those," meaning that during the determiner's emergence in speech the speaker does not just "commit" to a future formulation, to be "restored" after a hiatus as Clark and Wasow (1998) suggest: instead, she is already *fulfilling* her commitment to describe the referent. Though it may well be that her word repetition is due to a pressure to produce undisrupted syntactic constituents, if we ignored her gesture we might easily be led to the false conclusion that the abortive, pre-hiatus determiner is just a placeholder.

When the resumption occurs, the speaker's left hand stays in place in a continuation of the preceding hold, but her right hand switches to an abstract deictic gesture that fits the demonstrative linguistic item (see images [c1] to [c2], and notice that the main stroke of this deictic falls on "headsets," which is another case of logically associated content synchronizing in a fashion that would not be possible in speech alone—in fact it is a mirror of the previous combination, with the demonstrative now appearing in gesture while the noun appears in speech, though the 'headsets' gesture also persists in the other hand).[7] Next, although the gesture shown in image [d] looks rather similar to the configuration shown in image [b], it is actually quite distinct: it is an enactment of the usually one-handed action of gripping a bulky hand radio (a "walkie-talkie"), which contrasts with the earlier gesture's pinching of the ear-pieces of a head-mounted communication device (this was before the era of cellphones with a single Bluetooth earpiece). The fact that the gripping gesture of image [d] is done symmetrically, with both hands, may be a physical inheritance from the previous gesture's two-handedness, and underscores the fact that any gesture begins with and is influenced by the body's existing configuration as its starting point. Alterations in a gesture's starting point will require effortful change, especially when in the middle of a single rhythmic pulse of utterance (Tuite 1993; see

[7] This maintenance of an earlier gesture with one hand, while the other hand initiates a new gesture, is an instance of a gesture 'buoy', a phenomenon discussed in Chapter 2 (beginning on p. 57). More specifically, it is also an instance of Enfield's (2004, 2009) "symmetry-dominance construction," because it begins with a two-handed symmetrical gesture, followed by a new gesture performed with one hand while the other hand continues to hold.

also Loehr 2007), or “tone unit” (Kendon 1980). In image [e], the left hand has dropped to rest, but the right hand uses an unchanged configuration of the arm and wrist for its new gesture.

The question of speed and robustness of listener uptake

An objection could be raised regarding all of my claims of utility for the listener: can a listener really integrate these gestural cues, available at the start of constituents, fast enough to benefit from them? Might a listener instead wait until the linguistic constituent is complete before drawing any conclusions from its combination with gesture? Regarding the second question, if we follow Levinson’s (2000, pp. 27-35) tactic of an “argument by design,” it seems far more likely that listeners would use any and all sources of reliable information to make inferences and constantly update their mental simulations of a speaker’s intended message, as rapidly as possible. There are obvious advantages to knowing another’s next likely move earlier rather than later, and this information can come from any kind of sign, including non-linguistic ones.

But this still doesn’t answer the question of how quickly a listener can make use of perceived information of various kinds, and how that access really affects the dynamic course of utterance comprehension. As I mentioned in Chapter 2 (p. 48), neurocognitive studies have yielded suggestive results: Kelly et al. (2004) found that gestures about copresent objects, observed just before speech about those objects, yielded robust differences in neural response depending on the congruency versus incongruency of the gestures with speech. Wu and Coulson (2007) replicated those results in a study that lacked copresent object referents, obtaining the effect from iconic gesture alone. The differences in neural response were evident as early as 300 milliseconds after the word prompt appeared, and the effect was argued to indicate that information from gesture and speech begins to be integrated at the earliest stages of utterance comprehension. Furthermore, Özyürek et al. (2007) and Willems et al. (2007) found similar effects during comprehension of naturalistic utterances in which speech and gesture were fully synchronized, lending further support to these claims.

What are some of the functional results of a rapid integration of speech and gesture? In Example 3.1 above and others, I have shown the plausibility of visual information from gesture combining with incremental linguistic information to constrain reference at the very start of a constituent, before the linear segmentation of the linguistic portion is complete. This is very difficult to measure directly in natural settings where spontaneous gestures are observed, because directly testing a listener’s immediate uptake would destroy the naturalistic setting. Gullberg and Kita (2009) showed that a listener’s gaze fixations toward a speaker’s gestures (by a subject observing a video of a previously recorded face-to-face story narration) are not a reliable indicator of whether uptake of unique gestural information is taking place. However, this study faced the same obstacle of not being able to measure immediate uptake as it occurred: although they could measure gaze fixations continuously via a fairly unobtrusive eye-tracker, they had to wait until after the trials were complete to test the accuracy of each subject’s memory of gesturally presented information. They did not, therefore, test whether listeners' gaze fixations affected their immediate processing of utterances as they were perceived.

A number of other eye-tracking studies, while not directly investigating the role of gesture, still show clearly that visual, non-linguistic information about possible referents integrates rapidly with the incrementally updating information from language. This integration results in listeners immediately constraining their expectations of which referents are most likely, at least as far as can be gauged by preferential gaze fixations toward certain objects.[8] First of all, Chambers et al. (2002) showed that, when hearing a sentence about a set of objects, listeners take the visually available physical properties of the objects into account to rapidly arrive at preferential fixations, to the extent this is possible based on the information granted by each linguistic item as it is heard. When hearing "put the whistle inside the can," subjects were already fixating on the can container by the end of the word "inside," when there was only one container available in the set of possible referents. If the scene included more than one container, subjects did not preferentially fixate on the intended referent until hearing "can." But when the preposition was "below" instead of "inside," all of the objects in the scene were viable candidates for having something placed under them, and subjects in both conditions waited until hearing the noun referent before preferentially fixating on it. Additionally, when there was more than one of the same container present, but in differing sizes, subjects were able to rapidly fixate the intended referent when only one of them was the right size to fit the object in question, but were less successful when more than one of them was physically viable. The experiments by Chambers et al. (2002) deal with prepositions and their objects, but they also mention work by Altmann and Kamide (1999), in which verbs and their possible objects show similar effects. Upon hearing "the boy will eat the cake," subjects began fixating on the only edible item in the scene immediately upon hearing "eat," and this effect disappeared when the verb was replaced by the more broadly applicable "touch."

These studies' findings might in some ways seem difficult to compare to my claims regarding gestural cues, because they involve visual access to a scene, rather than fleeting access to temporary gestures. However, two points help bridge this gap: first of all, gestural holds are more like stationary objects than are other kinds of gestural cues, and they have also been shown to draw visual fixation by the listener, an effect that is strengthened when the speaker also fixates on them (Gullberg and Holmqvist 1999, 2006; Gullberg and Kita 2009).[9] Second, some of my examples involve deictic reference to aspects of the model village, which makes the village scene analogous to a very complex version of the visual displays used by authors such as Chambers et al. (2002). In these cases, the question becomes: how might gestural cues, *combined with* the features of the model, influence listeners as they are constraining reference at the start of utterances? I tackle this question shortly in Example 3.2.

[8] See Eberhard et al. (1995) for a discussion of eye-tracking methodology, as used for studying incremental processing as well as contextual dependence during language comprehension.

[9] This is consistent with Streeck's (1993, p. 286) characterization of gaze directed at one's own gesture as being "a marker of the communicative relevance of [the] gesture."

Disfluency as a cue used instantly by listeners to constrain impending reference

Arnold et al. (2003, 2004) revealed a phenomenon related to those just discussed, one which is especially relevant to the theme of this chapter and Chapter 2. Non-invasive eye-tracking of the listener was once again the means of measurement. The authors showed that fluent renditions of the definite article plus a noun (as in “the candle”) caused listeners to start fixating preferentially on referents that had already been mentioned, while disfluent production (as in “thee, uh, candle”) caused listeners to start fixating preferentially on referents that had not yet been mentioned. When the *disfluent* version was used with a referent that was ‘given information’ instead of new, the eye-tracking measurements showed that listeners reversed a starting strategy of fixating on the new referent, and began fixating on the correct, already-mentioned referent by around 400-500 milliseconds after the noun’s onset. Similarly, when the *fluent* version was used with a referent that was new instead of given, listeners had to reverse their starting strategy of fixating on the already-mentioned referent, to begin fixating on the new referent instead.

What these studies show is that listeners respond not just to content as it is incrementally perceived, but also to the way it is delivered. Besides initiating an utterance and showing a speaker’s intention to continue, at least some of the preliminary commitments discussed by Clark and Wasow (1998) can evidently reveal to the listener that the resumptive utterance will most likely be about something new, instead of something already mentioned. Information from gesture could then help pinpoint what this new referent will be.

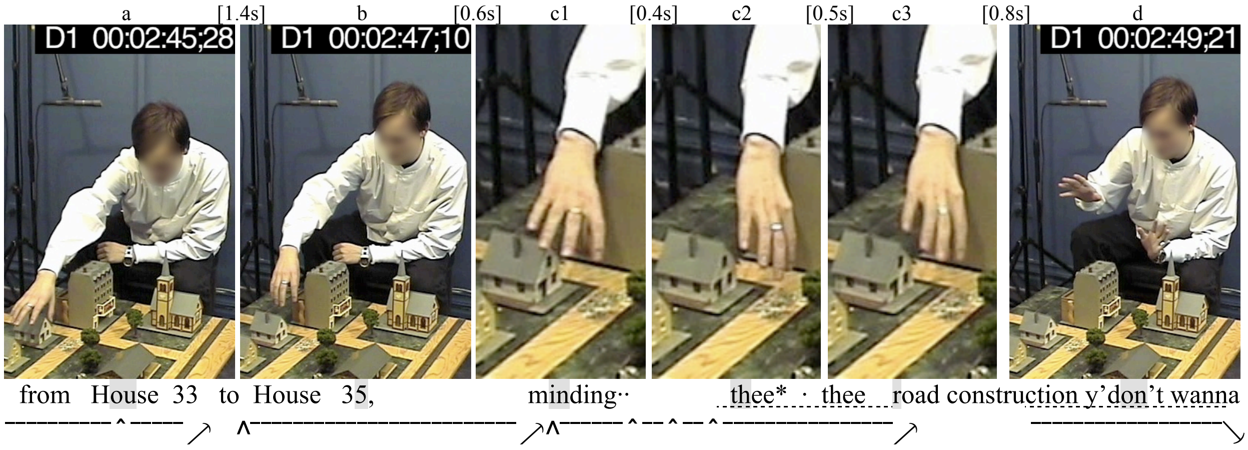
Examples 3.2 and 3.3 — deictically constraining the range of likely new referents

The first example below demonstrates a scenario resembling Arnold et al.’s (2003, 2004) studies: it includes a hiatus after the word “thee,” supported by a deictic hold precisely pinpointing a specific feature of the model village that has not yet been mentioned. The speaker is fixating on his own gesture and the referent, which increases the likelihood of the listener attending to them (Gullberg and Kita 2009). Listeners are evidently adapted to react to the fluency of the start of a constituent by searching for likely referents, and in a setting with a great many possible referents (such as when viewing the model village), they should also be able to make use of the speaker’s gestural cues.

As with Example 3.1, the speaker’s gesture in this example is already in place and providing content by the time the speech suspension takes place following the determiner. The gesture is a close deictic hold pinpointing a set of objects merely an inch away, which should handle most of the work of constraining reference among the many features of the model. But given Arnold et al.’s (2003, 2004) findings, the suspended determiner may additionally reveal to the addressee that the referent is something new, as opposed to something located at almost the same position but already mentioned (such as House 35 itself). Throughout the passage, the listener is staring intently at the part of the model being indicated. She will not know exactly what the ornamental fishtank pebbles glued to the model are meant to represent until her partner completes his utterance, but she should guess that they are the object of “thee.” (The pebbles are somewhat difficult to make out in the snapshots; they are the mottled smudge in the light-colored ‘road’, visible just to the right of the house in each of the three zoomed-in images.)

(3.2) **A**: And then* uh·· ··· So we’re going · ta· go ·· to this house ·· #
and then we’re going· to ··· bring· these* ·· our assistants from H[a]ouse 33 to House 3[b]5,
m[c1]inding·· [c2]thee* · thee [c3]road construction y’[d]don’t wanna fall in there or anything.
I can’t afford to l* ·· # to have anybody injured on this mission.

a-d: *Throughout, Participant B (not shown) is leaning forward, right elbow on her right knee and holding her chin with RH, gazing intently at the part of model being indicated by Participant A. As he is saying “thirty-three to house thirty-five,” she makes several small nods.*
a-b: *Participant A was holding a loose RH deictic over House 33, then holds over the nearer House 35.*
b-c3: *His hand jerks to his left to hover over the fishtank pebbles glued to the wooden board (image* [c1] *occurs at* [02:47;28]*), after which his fingertips perform three extremely small and quick movements toward the pebbles. His whole hand then begins waggling as he says “thee” (image* [c2] *occurs at* [02:48;11] *and image* [c3] *occurs at* [02:48;26]*).*
c3-d: *He retracts his forearm without leaning back, and BH raise at wrists to waggle with palms out.*



Finally, in Example 3.3 below (which I will show more of in Chapter 5, as Example 5.2), Participant B repeats the determiner “the” after a hiatus spanned by her already-deployed deictic hold, but it involves a revision of the utterance rather than a simple repeat: she inserts an interrogative phrase (“what is that”) to revise the impending utterance into a query about the identity of the referent. By the time she utters the preposition “in” (image [b2]), her partner has already turned her head to look at the target of the deictic hold, an action which occurs about 400 milliseconds after Participant B’s deictic gesture first takes form (the shift in Participant A’s head angle is virtually impossible to see in the snapshots, but it is a very slight turn to her left between images [b1] and [b2]). As we saw above in the work by Chambers et al. (2002), upon hearing the preposition the listener will very likely be searching for plausible referents where people in the town could “work in," and the speaker's hesitation makes it more likely that the referent will be something not recently mentioned. Thus, by the time her partner finishes saying "what is that the park," Participant A has had a great deal of time since the start of the continuously maintained deictic to make use of information from gesture, speech, and the model. She will have begun processing the most likely referents, to be ready to respond to her partner's queries.

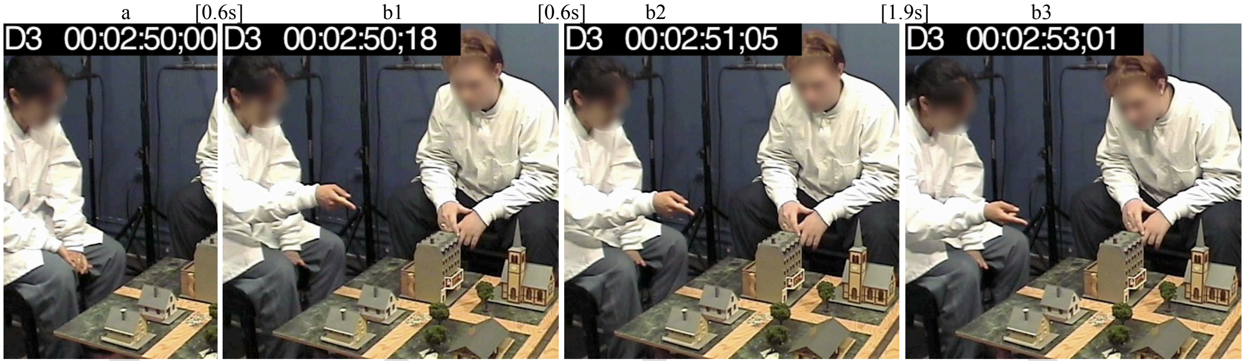
(3.3) **B**: So in the directions did you receive · information about any other people in this town?

```
         a                  b1          b2
  ⌈A:   #   ·  No.                                                 ⌉
  ⌊B:  Like · does nobody work · in · the ·· what is that the park?  ···  ⌋

                   b3
   A:  That's · that's a train station?
```

a-b1: *While Participant B gazes at the model, her RH rises from her lap to form an index finger deictic hold that targets the trees and building visible at the lower right of the images.*

b1-b2: *Less than 400 milliseconds after her partner's deictic takes form, Participant A turns her head almost imperceptibly to her left to attend to the target of her partner's gesture.*

b2-b3: *As her partner maintains the hold, Participant A leans forward toward the mutually attended referent and responds while raising her eyebrows.*



```
[A:      #     ·    No.                                               That's · that's a train station? ]
[B:     Like   ·   does    nobody    work   ·   in · the · what is that the park? ···                   ]
(B's RH)  ___________ ↗      ^ ---------------------------------------------------------------------------
```

Kita (1993) found that pure repetitions in speech, in which no repair is evident, were much more likely than repairs to coincide with a continuation of the gesture phrase. However, he still found that 39% of repairs coincided with continuation of the gesture phrase, and in 19 of the 21 such cases in his survey there was "retention of an image" across the repair—that is, a potentially meaningful gestural cue was maintained across the repair and into the resumption (Kita 1993, p. 74). Kita classifies these gesture-retaining repairs as "surface repairs," because much of the utterance, including the gestural content, is still valid and it is only some element of the linguistic stream that needs to be revised. Although the revision in Example 3.3 is not necessarily a repair in the sense of fixing an “error” in the linguistic stream, it is similar to Kita’s examples in that the maintained gesture retains its validity throughout the adjustment in speech.

The example also illustrates the lack of a required termination of a post-stroke hold to coincide with (and mark) the end of the unpacking of a gesture’s co-emergent linguistic material. This requirement is part of the traditional account of post-stroke holds that Kita (1990, 1993) and McNeill (1992, 2000, 2005) have emphasized, but it is challenged by examples like these, and by the phenomenon of gesture ‘buoys’ illustrated in Chapter 2 and by Enfield (2004, 2009). The termination of a gesture may instead be under high-level control, as De Ruiter (2000) suggests. It certainly cannot be the case that a gesture remains in play until the precise moment that its associated “lexical affiliate” is heard by an “auditory monitor,” as Krauss, Chen, and Gottesman (2000) believe. Even if we could reliably identify a discrete “lexical affiliate” from the speech stream (and we cannot, as De Ruiter and many others have stressed), the idea that arrival at such an item triggers the termination of the gesture is simply inconsistent with observed behavior, including all the observations leading to the models of Kita and McNeill and earlier work such as that by Kendon (1980), who showed that many gestures end in time with the full completion of co-timed “tone units” of speech.

As Example 3.3 and others show, a speaker can evidently choose to keep a gesture in play through the end of an utterance and even into an interlocutor’s turn, when this might serve some purpose. In this case, the persistent deictic continues targeting a referent that the interlocutor will be speaking about, assuming she accedes to her partner’s request for information. Maintenance of gesture holds across turn boundaries may therefore serve conversational goals. I will return to this theme in Chapter 5.

The question of speech-gesture synchrony

McNeill’s seminal 1985 paper built on Kendon’s (1972, 1980) work promoting a common origin for speech and gesture, a line of thinking which later led to the development of Growth Point Theory (McNeill 1992). Among the supporting lines of evidence given by McNeill (1985) was the claim that speech and gesture are closely synchronized in time. Even from this early date, it was never a claim of just the simple synchrony between discrete gesture strokes and single words of speech, which some utterances certainly exhibit. McNeill (1985, p. 353) also mentions “sentence and gesture synchrony,” very much related to Kendon’s (1972) discussions of gestures and speech grouping together into the prosodic phrases (“tone units” in Kendon 1980) that make up “locutions.” So for both Kendon and McNeill, early statements regarding speech-gesture synchrony were first of all about gestures synchronizing with linguistic units of one size or another, and within this framework McNeill’s production model also ascribed importance to more precise moments of synchronization such as that between discrete strokes and single morphemes.

What should be evident is that speech-gesture synchrony is unavoidably complex. Most gesture strokes occur within a larger process in which the onset of movement is significantly earlier than the stroke. One reason for this is simply the inescapable fact of the inertia of our mass-bearing articulators: it takes time for the arms, hands, etc. to move from rest position to the shapes or spatial arrangements that they take on during utterances, such as when gesture strokes occur. If the onset of movement marks the very beginning of the co-emergence toward gesture and speech of a new expressive thought (the onset of what would be termed the “Growth Point”), then the stroke marks the peak of expressivity in both channels. As McNeill (1989) emphasized in his reply to Butterworth and Hadar’s (1989) critique, many findings seemingly at odds with his claims of synchrony may simply boil down to the issue of making mismatched timing comparisons.

An example of such a mismatch would be a comparison of the timing of a single important word in the speech stream to the first onset of gesture movement (the start of the preparation phase). We see this in the work of Morrel-Samuels and Krauss (1992), who found that gestures last longer, and also anticipate their lexical affiliate by a greater amount of time, the less familiar the word. They in fact agreed with McNeill’s stance that speech and gesture are not produced by independent modules, but it is noteworthy that they characterized their findings as showing “familiarity’s systematic relations with *gesture-speech asynchrony*” (p. 615, emphasis mine). This so-called “asynchrony” does not actually conflict in any way with the possibility of close synchrony between gesture strokes and the authors’ reported lexical affiliates, because they defined gesture onset as any movement faster than 1.5 inches per 0.8 seconds (p. 617), and the concept of the stroke was not remarked on anywhere in their paper. In other words, their findings are consistent with a prolongation of the preparation phase leading up to the stroke, when the linguistic material is more difficult to arrive at due to being less familiar.

One of the reasons for the confusion may have been that McNeill (1985) did not make clear the distinction between gesture phases first proposed by Kendon (1980). The word ‘stroke’ was only mentioned in one diagram by McNeill (1985, p. 361), meaning that the paper was rather vague in terms of which parts of gesture were claimed to synchronize with units of speech. This shortcoming was forcefully rectified in his 1989 reply to Butterworth and Hadar (1989), but in the process he appeared to miss one of their criticisms: they pointed out that some authors, notably Schegloff (1984), had claimed that it is not just gesture onsets, but also gesture “thrusts” and “acmes” which occur significantly earlier than the closest “lexical affiliate” in the speech stream. Schegloff’s stance remains highly influential; for example, Levinson (2006) believes there remains a “binding problem” between speech and gesture, in part because of their apparent lack of temporal binding: since they are not well synchronized, how do we know to associate them? But it turns out that much of this binding problem dissolves when we discover that the entire notion of Schegloff’s “lexical affiliate” may itself be invalid. As De Ruiter (2000) argues, there are many utterances which simply lack any single word or phrase that clearly matches the concurrent gesture most closely. And although it is possible for subjects to choose a single word from transcript which they believe most closely matches the meaning of a gesture, as was the procedure of Morrel-Samuels and Krauss (1992), this does not mean that it is the only word or phrase interpretable as being associated with the gesture by the speaker, nor does it mean it was necessarily part of the same utterance pulse by the speaker.

This last point is crucial: speech and gesture frequently present *complementary* content such that a complete utterance may present highly associable, coherent information even though none of the words closely reproduce the content of the gesture. But a speaker may then elaborate with speech that reproduces some of the content of the earlier gesture. This easily creates the impression of a binding problem between the later speech and the earlier gesture, but it is illusory: the emergence in speech of content closely matching the earlier gesture does not in any way eliminate the associability of the simultaneous speech and gesture in the initial utterance pulse, in which complementary but *congruent* information was presented together in the two modalities. There is a partially causal relationship between all sequential utterance contributions, and this once again does not in any way detract from the associability of multimodal contributions that are simultaneous. Cross-modal priming may exist, but as a sequential relationship such as we find in priming phenomena in general. In this fashion, we can simultaneously accept the McNeill/Kendon stance regarding co-expressive speech and gesture performed together, and also accept the position of Krauss et al. (2000) that gesture can affect subsequent lexical retrieval in speech.

Some examples from Duncan (1996) illustrate the point: consider an utterance (p. 85, translated from Mandarin in its original syntactic order) consisting of "this cat TOPIC MARKER on the big street rolls around" performed with a gesture, timed with "this cat," that expresses a circular 'rolling around' motion. The fact that "rolls around" also appears in speech within a second or two does not in any way detract from the congruency of the separate elements at the start of the utterance, in which a figure entity (the cat) is given in speech simultaneously with a gesture expressing the motion of the figure. As Duncan points out, the "further unpacking into speech" of the gesture's content is not something that always occurs. In another example (p. 84), an utterance consisting of "There's an organ grinder in the street" is accompanied by an iterative iconic enactment of cranking a street organ by the handle. In this case, the motion of the organ grinder's action, as well as its progressive aspect, is inferable from gesture but appears nowhere in speech. And yet the intended association of the elements from the separate modalities remains unambiguous. If the speaker had gone on to elaborate with additional speech closely replicating the content of the gesture, such as the phrase "cranking on his organ," would we suddenly be justified in claiming asynchrony of speech and gesture?

Another potentially problematic case might be a gesture spanning multiple clauses: if a stroke times closely with one phrase of speech, does a lingering gestural artifact, such as an extended post-stroke hold, remain "synchronized" with multiple clauses of additional speech? McCullough (2005, p. 630) suggests there is no problem provided we allow for scope-level coordination, rather than require speech and gesture to be co-expressive at the same level: "the requirement that they be coordinated is more akin to the coordination of speech elements in a scope relation—they are in fact related above the level … of the more strictly co-expressive clusters that David McNeill [1992] has focused on." This scope-level coordination is basically another form of speech-gesture complementarity.

The upshot of all this is that the entire debate on speech-gesture synchrony may boil down to a misunderstanding of what a gesture must occur with in order to form a synchronized composite. I have found no examples in my corpus of gestures occurring with 'incongruent' speech to form illogical or uninterpretable combinations. We can allow for the influence of elements of any modality (e.g., gesture and/or speech, and physical artifacts such as the model village) on subsequent utterance formation and interpretation, while still believing in the co-expressiveness of those elements that are simultaneous.

Example 3.4 — speech-gesture synchrony amid apparent asynchrony

Consider the example below, which I had originally selected as a case that seemed to strongly challenge McNeill's claims of speech-gesture synchrony. The transcript is a direct continuation of Example 3.1 above: Participant B has just finished saying "but… those… those headsets not walkie-talkies, walkie-talkies are too noisy," to which her partner responds as shown at the start of the transcript below. In Participant B's next utterance, her gesture enacts an iconic (metaphoric) gesture for 'sound emission' bursting from her mouth or from a walkie-talkie held there. The trouble is that, if we are looking for the speech whose content replicates the gestural content most closely, we will end up with the phrase "talking into walkie-talkies is noisy," which is several seconds later and in a completely separate tone unit, following a breath. At that point, she has long since stopped gesturing. Her hands are already at rest beginning with the moment of image [c]:

(3.4) **A**: oka*you'll need* and you'll need ·· the headsets so you h*have your hands free.

B: Yeah. ····

a b1 b2 c d

Well and the wombat* · we don't want the wombats to hear us coming. #

Y'know when · talking ·· into ·· walkie-talkies is noisy.

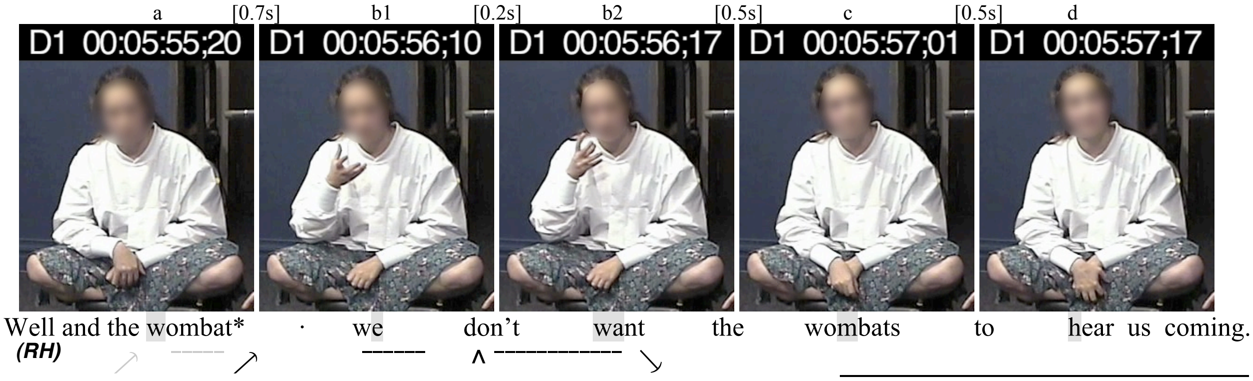
a: *RH has begun slowly rising from lap but hesitates in place when it is just above her other hand.*

a-b1: *From the end of the word "wombat," RH rises quickly to hold in a loose claw at her mouth.*

b1-b2: *Timed with the word "don't," she executes a stroke at* [56;13] *consisting of a sudden twist of the wrist as her fingers clench slightly and release. Thus far, she has been looking at the model.*

b2-c: *RH drops back to rest as she shifts her gaze to her partner's face.*

d: *The phrase "hear us coming" is performed with markedly raised eyebrows and several nods of her head. She continues to elaborate in speech, and her gaze drops back to the model in the pause before the word "talking." She performs no additional hand gestures, but leans back and adjusts her legs slightly as she pronounces "walkie-talkies is noisy" with clear utterance-final intonation.*



The word that the stroke actually synchronizes with, meanwhile, is "don't." But because of the existence of later speech which describes the content of the gesture more closely, a study such as that of Morrel-Samuels and Krauss (1992) would consider this a case of "gesture-speech asynchrony." For Levinson as well, this example would illustrate the "binding problem" (Levinson 2006) that he has argued remains unsolved in our theories of multimodal communication. And Krauss et al. (2000) would presumably argue that the gesture is facilitative of the speaker's lexical retrieval of the later linguistic items (though this is a rather dubious prospect here, given that she has already mentioned in recent speech that "walkie-talkies are too noisy," at the end of the final utterance of Example 3.1).

In spite of these difficulties, McNeill (personal communication) does not, in fact, find this example problematic. At the level of the "tone unit" (Kendon 1980) or utterance pulse, whose speech here consists of "we don't want the wombats to hear us coming," the gesture expresses completely congruent content: it demonstrates part of the event that would lead to the wombats "hearing us coming." Furthermore, at a closer level of synchronization, the gesture stroke coinciding with "don't" is still congruent, because the sound emission expresses exactly 'what we *don't* want'. As I have stated, McNeill's later writings (e.g., McNeill 2000) may have overemphasized the degree to which the gesture should *end* in time with the end of the associated tone unit, which in this case would include the phrase "hear us coming." But a gesture which ends before, or after, the end of an associated tone unit is no less associated with it. It suffices that there is *some* period of simultaneity.

For there to be a true "binding problem" between multimodal signals that are removed from each other in time, we would need to find cases of intermodal association which are decidedly *not* congruent in terms of what is occurring simultaneously, and *only* associable across a significant gap in time. Since I have not found this, I believe that at the level of moment-to-moment, logical unity of expression, it may simply be enough to follow the operating heuristic (see Levinson 2000) that co-timed elements are associated in some fashion (determining the details of this association is a separate question). Meanwhile, across significant gaps of time there may be additional intermodal associations which seem even closer, in terms of matching content, than the associations between what is simultaneous. But this may be the same associability and causality we normally find between different points in the output of a single modality: there is certainly priming within a single modality such as the stream of speech, and there is no reason to deny that intermodal priming would occur as well, such as the "cross-modal priming" claimed by Krauss et al. (2000, p. 269), though we need not agree that the primary purpose of gesture is to bring about such priming.

In fact, it is precisely because of a lack of a binding problem between speech and gesture that I believe listeners are able to make immediate use of complementarily presented gestural information, including in cases of speech disfluency where the exactly simultaneous speech's content (or lack thereof) may become less redundant with gesture and 'more complementary' than in fluent utterances. In the oft-cited example from Kendon (1980, pp. 219-220), a speaker is discussing Billy Wilder's *Some Like It Hot* and says, "they wheel a big *tab*le in/ with a big (pause) *cake* on it." During the pause, the speaker performs a gesture indicating the shape of a cake, and while there is no ambiguity about the gesture's co-expressivity with the tone unit (consisting of "with a big (pause) *cake* on it"), it remains the case that the listener has access to gestural cues about the cake referent before the speech directly naming it. Stroke phases evidently do not always time precisely with a stressed syllable of the co-expressive linguistic stream; instead, as Clark (1996, p. 177) suggests in discussing this example, the speaker may perform her stroke "where she projected *cake* would occur." A similar event was illustrated in Example 2.3 in Chapter 2 above, when the speaker performed a 'corner' shaped gesture during a silent pause, prompting her partner to suggest words that would complete the linguistic stream. Disfluent speech often increases the apparent disjunction in the timing of speech and gesture, and examples like these may therefore have contributed to the belief that speech and gesture are frequently "asynchronous."

Early on, McNeill (1985, p. 361) did acknowledge that gestural content sometimes emerges earlier than associated linguistic content: "There exist anticipations where the concept revealed in the gesture becomes available before the sentence can grammatically make use of the linguistic item that signifies the concept." This was before our more recent understanding of speech-gesture complementarity was worked out, and he characterized such cases as explicitly *not* synchronized, in spite of his earlier mention (p. 353) of broader "sentence and gesture synchrony." His example in that paper seems roughly comparable to Example 3.4 above (which he today finds quite synchronized). In the passage, a deictic gesture for the location of a character coincides with the words "flashing back," during the utterance "they keep flashing back to Alice just sitting there." At the time, McNeill noted that the timing revealed that the "incident of thinking" of Alice's location "anticipated by four words the verbal reference to the location," which he singled out as the word "there." But today, I believe McNeill would not be troubled by the gestural anticipation of that specific word—in fact, as I have discussed with the organ grinder example from Duncan (1996), there is no requirement that the word "there" appear in speech at all. A deictic gesture coupled with just "they keep flashing back" could already be perfectly associated to represent *where* "they keep flashing back" *to.* Gestures can combine smoothly with many of the words of an utterance, not just those items that seem to most closely replicate their content. I see no reason to privilege redundancy over complementarity, in judging what is associated.

In these chapters I have emphasized that disfluent speech can bring into sharp relief the phenomenon of gestural content being interpretable by a listener early on, while much of the intended simultaneous speech is still missing. If the gesture remains available as a *hold,* it may then *also* later synchronize with speech that describes or names content similar to that of the gesture. Recall Example 3.2 above (p. 80), in which the gesture is indicating the 'road construction' pebbles in the model village. The gesture appears first with a stroke timed with the verb "minding" (i.e., the listener could reasonably assume the gesture is indicating the thing that will be minded), and then remains as a hold during two instances of the determiner "thee" (confirming that the gesture is indicating a noun entity which is the object of the determiner, in a phrase which is the object of "minding"), and lastly it is still held (and shortly thereafter released) when the noun itself finally appears in speech. It seems entirely wrong to characterize such cases as "asynchronous" merely because a gesture spends much of its existence, or even all of its existence, in association with linguistic items other than those that precisely name the referent.

Recurring speech-gesture pairings in utterance comprehension

In McNeill's strong version of synchrony, gestures include a stroke phase that is closely timed with a stressed syllable of a co-expressive linguistic item, and the stroke is characterized as "the gesture phase with meaning" (McNeill 2005, p. 32), and the phase that "carries" its "content" (McNeill 1992, p. 84). By extension, preparatory phases (the preparation and pre-stroke hold, if present) are not deemed expressive, and McNeill (2000) makes explicit the idea that the contrast between preparatory phases and stroke are a useful way for speakers to indicate which elements in their speech are *not* part of the co-originating Growth Point with the gesture. One problem with his stance, as I have shown, is that meaningful parts of gestures appear to be capable of emerging 'on time' even when speech is delayed due to disfluency—gestures do not always wait for the maximally co-referential elements of the utterances they associate with. But another problem is that, even with examples that are completely consistent with McNeill's stances on the timing of the stroke, a model focused only on production appears to ignore the highly significant information available to the listener during the preparatory phases. Even if the onset of the preparatory phases marks only the start of the emergence of a speaker's Growth Point, and even if the speech during those preparatory phases is not yet part of that Growth Point, the gesture's inchoate pre-stroke form can still influence a listener's utterance comprehension. It could even be argued to be 'expressive' in the sense that it presages some of the content that the speaker is about to express.

I have emphasized that in disfluent speech-gesture combinations, listeners can have access to gestural cues that reveal or at least thoroughly contextualize the speech that finally arrives. This is sometimes evident before any words escape the lips: recall Example 2.8 from Chapter 2, in which the speaker inhaled in preparation for speech and also renewed the muscle tension of her deictic gesture, which she had been keeping in a loose hold during a temporary digression (see p. 56). Although she breaks into laughter and has to pause again before actually resuming her former discussion, her *bodily* resumption just before the laughter, including the reestablished deictic focus on a particular target, is unambiguous. Kendon (1972, pp. 207-208) has noted that these preparatory movements may serve as advance warning of turn-takings, as "floor-apportionment" signals (Kendon 1967), and that the severity or type of the bodily shift could signal the size or type of the speech unit to follow. I now wish to show that even with completely ordinary fluent utterances, with unproblematically timed gesture strokes, the preparatory phases can sometimes be so revealing that a listener could easily guess the *content* of the impending utterance, and even correctly guess the speaker's choice of words. This remarkable state of affairs arises because of a particular phenomenon that often occurs when speakers refer to the same referent multiple times within a single discourse.

Just as with any utterance, the first mention of a recurring referent may include some novel gesture that expresses something about it. It happens that when mentioning the referent again later in the discourse, speakers have a tendency to use both a similar wording and a similar gesture. There may be important psychological reasons for this, in which previous lexical choices prime themselves to become favored choices later, and previously performed gestures similarly prime themselves to reappear. In any case, the phenomenon is quite common, and McNeill refers to these recurrences of gestural form as ‘catchments’ (see McNeill 2005, pp. 116-118, for definitions). The key here is that the recurring speech-gesture pair demonstrates a form of on-the-fly, ephemeral conventionalization. Through repetition, the form of the gesture stops being merely iconic (for example), because it becomes identifiable with the full speech-gesture complex, including the recurring lexical choices. As McCullough (2005, p. 713) puts it, the catchment becomes a kind of “nonce Saussurian symbol.” Slowly, and whether the speaker intends it or not, a listener can begin recognizing the gesture on its own as a kind of *label* for the referent, rather than a token gestural representation of some of the referent’s features. This is the same process discussed by Kendon (2004, pp. 307-309) as underpinning transitions from novel gestures to conventionalized signs, which eventually include a drastic reduction in complexity until the sign is minimally contrastive and quite different from its original detailed enactment, depiction, or modeling. A brief slice of the process he describes takes place in routine conversations: as a speaker begins to use a recurring gesture more as a label, it eventually tends to reduce in size and complexity.

This nonce typification or conventionalization through recurrence in a discourse is a mechanism relying on the ‘retrospective’ relationship of each recurrence with previous instances. Those instances must remain accessible in some fashion (i.e., in memory) in order for the recurrence to be noted, an effect which is presumably possible because of their recency in the same discourse. Unlike the fashion in which an extended hold can span part of a discourse, a catchment emerging as a discontinuous recurrence is a gesture which undergoes a full set of gesture phases each time—that is, it must be enacted with every reappearance, and cannot simply be ‘held over’ from earlier moments. Because of this repeated reenactment of paired content from speech and gesture, interlocutors can become trained to expect particular speech each time the catchment’s distinctive form becomes recognizable.

Example 3.5 — catchments may trigger early prediction of full utterances

The following five-part example illustrates the development of one such ‘catchment’ which is distinctive enough in form that it eventually becomes quite recognizable even during the preparatory phases. In its first occurrence (3.5a below), the catchment is not yet well defined, but it begins taking form in images [b1] and [b2] as the speaker mentions “soundblasters” for the first time. The form of the catchment always involves the pinky, ring, and middle fingers of both hands becoming curled, with the index fingers extended and the thumbs raised as if ready to press some kind of trigger or button.

(3.5a) **A**: What* What do* What · prey on wombats? ·

What* What eat wombats? · What are their natural predators? ··

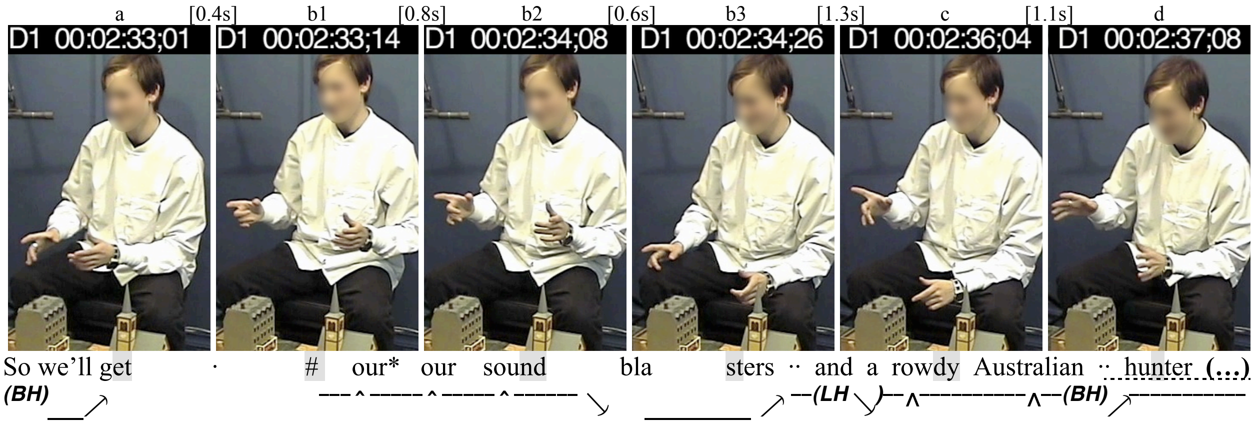
B: Umm··· cheetahs! · #

A: How ’bout row* r* ·· rowdy Australians? ··· Yes! #

[**A**: So we’re going to need to make a soundtrack of that to scare them out of the theater.]
[**B**: Okay.]

a b1 b2 b3 c d
A: So we’ll get · # our* our soundblasters ·· and a rowdy Australian ·· hunter soundtrack.

a-b1: *BH rise from lap with loose open palms facing each other, then three fingers of each hand curl in slightly, leaving index fingers extended, as wrists rotate to point thumbs upward.*
b1-b3: *This configuration is given three beats (or small strokes), then hands drop to lap while retaining this vague shape.*
b3-d: *BH rise again momentarily, then LH drops back to lap while RH extends index finger more emphatically for two strokes directed at his partner, after which LH rises again and BH open up to spread-fingered palms which waggle in place.*



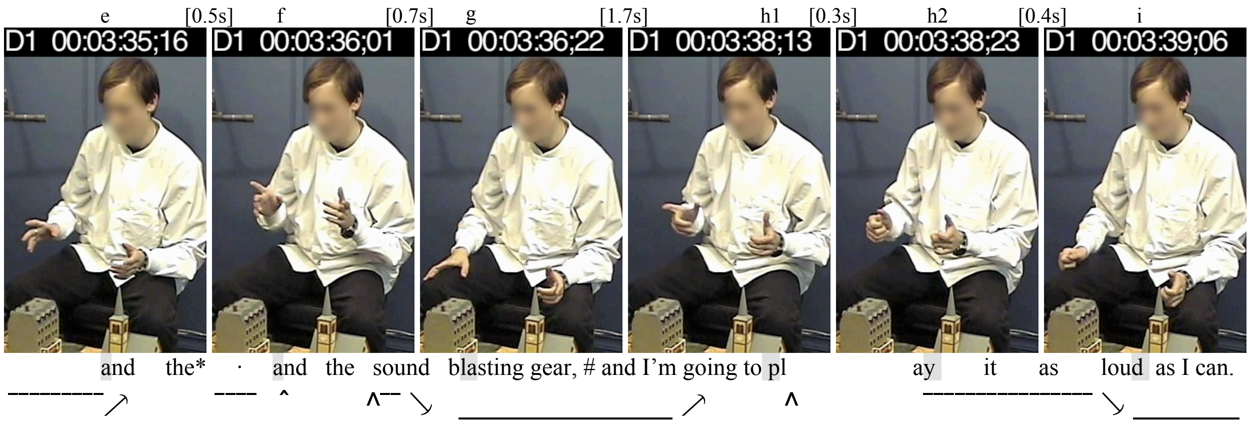
Images [b1] and [b2] show the very first, somewhat ill-defined emergence of the catchment handshape, and I would not claim that the listener will be able to predict much about the intended object of “So we’ll get our* our…” in the utterance as it leads up to image [b2], except that it may possibly be related to the recently mentioned goal of making a loud soundtrack to scare the wombats out of the theater.

However, in 3.5b below, the handshape appears very briefly (image [f]), once again associating with “soundblasting” gear, followed by an extremely clear and emphatic enactment of actually activating the gear, shown in images [h1] and [h2] (the difference in how clear and emphatic the stroke is here is not easy to show in snapshots, but it is quite apparent in the original video).

(3.5b) **A**: And meanwhile I’m going to go to the front with · whoever I have, ·
e f g h1 h2 i
and the* · and the soundblasting gear, # and I’m going to play it as loud as I can.

e-g: *BH have been holding with loose curled fingers, but at the onset of the first “and” the three lower fingers of each hand curl slightly as the arms raise (the index fingers remain extended). The arms pause momentarily before rocking up on the second “and,” after which the hands drop to lap but with a stroke-like emphasis at the start of “sound.”*

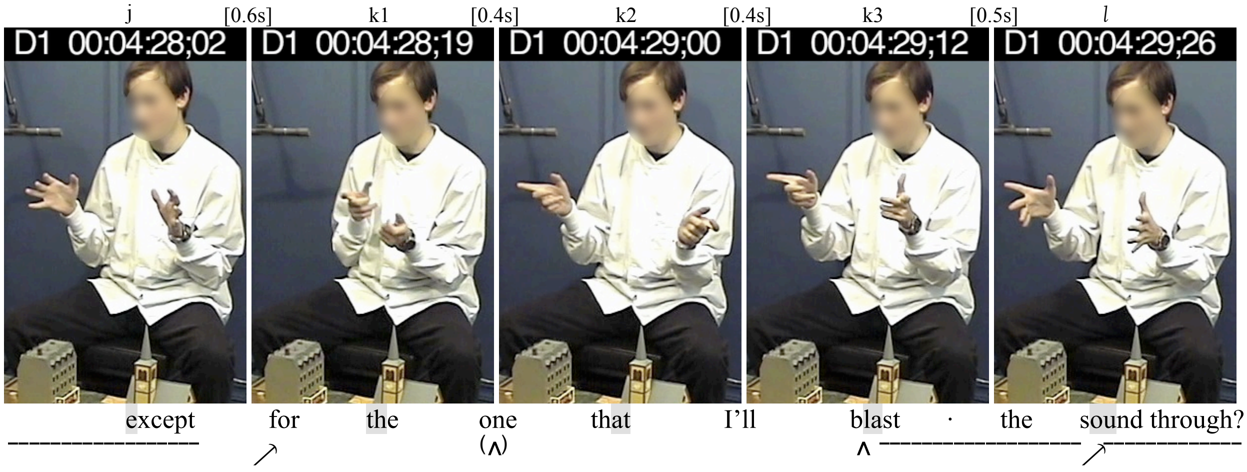
g-i *BH rise into symmetrical shapes that repeat the previous gesture but with much more definition: the thumbs are extended upwards, the index fingers are extended, and the other three fingers of each hand are braced together and curled into ‘c’ shapes as if gripping handles. BH jerk forward with a very sharp stroke (timed exactly with “play”) in which the fingers grip tightly (including the formerly extended index fingers), and the RH thumb squeezes down on the fist, while LH thumb stays extended. BH then drop to rest and retain this grip shape for 1 second.*


e [0.5s] f [0.7s] g [1.7s] h1 [0.3s] h2 [0.4s] i
D1 00:03:35;16
D1 00:03:36;01
D1 00:03:36;22
D1 00:03:38;13
D1 00:03:38;23
D1 00:03:39;06
and the* · and the sound blasting gear, # and I’m going to pl ay it as loud as I can.
↗ ˆ ^ ↘ ↗ ^ ↘

After the passage shown in 3.5b, I would argue that any further appearance of the handshape could be interpreted by the listener as a likely signal of impending mention of the soundblasting action. If the listener developed this expectation, she would correctly anticipate the content of the speaker’s utterance in each of the three subsequent instances of the catchment, at least insofar as an assumption that the utterance would include a mention of the speaker “blast(ing) (sound)” through the front doors of the theater. The emphatic gesture stroke in this passage occurs at [03:38;15], which is only 100-200 milliseconds after the handshape gains its unique form during the preparation phase (roughly from [03:38;10]), and this is too short a time for me to make any strong claim of a ‘presaging’ influence on the listener. But in subsequent instances of the catchment there is a great deal more time leading up to the stroke, during which a confluence of factors make the speaker’s mention of “blast(ing)” seem all but inevitable. This is evident in 3.5c below:

(3.5c) **A**: And so·· well actually, # I will ··· go into the theater. I won't be outside.
I'll go into · right where the theater is and I'll close · all of the doors, ·
j k1 k2 k3 *l*
except for the one that I'll blast · the sound through? #

j-k3: *BH have been holding with spread-fingered palms facing outward, then BH switch to the shape seen in earlier instances, with the lower three fingers curled and the index fingers and thumbs extended, while the forearms swing inward. Elbows stay stationary, but forearms then swing outward widely on "one" and reach their widest extent on "that," then come inward again for a sharp (but short) downward stroke on of the hands on "blast," which is followed by a hold.*
k3-*l*: *While BH maintain their position, the fingers open up into spread palms, which continue holding.*



This time, the catchment handshape comes into full form nearly one full second before the stroke coinciding with "blast," enough time for the speaker to flail his arms out and back loosely during the intervening speech (an action which may be a superimposed metanarrative gesture). This passage in 3.5c also occurs only 50 seconds after 3.5b, meaning there is an excellent chance that the speech-gesture pairing is still fresh in the listener's mind. In addition, the speech during the preparation effectively sets up the relevance of the doors to the theater, and one in particular which will be left open, where he will do something. At this point in the conversation the 'blasting' activity is the main plan that he has personally taken responsibility for performing at the theater, so the listener could predict this as one of his most likely referents from the preceding speech alone. With the addition of the gesture preparation, she would be all the more sure of it.

Next, in 3.5d, the preparation occurs during a silent pause. The form of the preparation could once again signal to the listener what the utterance will be about, except that it takes shape only about 200 milliseconds before the stroke. The preceding speech and gesture have once again identified 'here' as the front of the theater, where the speaker has located himself, which may be predictive. Mostly, 3.5d serves to demonstrate the variability in how long the preparatory phases last during the lead-up to each stroke; it isn't the case that every instance will reveal itself strongly ahead of time. It is, however, another reinforcing instance of the speech being associated with the gesture. With each recurrence, the ability of the gesture catchment to stand for the full speech-plus-gesture complex becomes stronger.

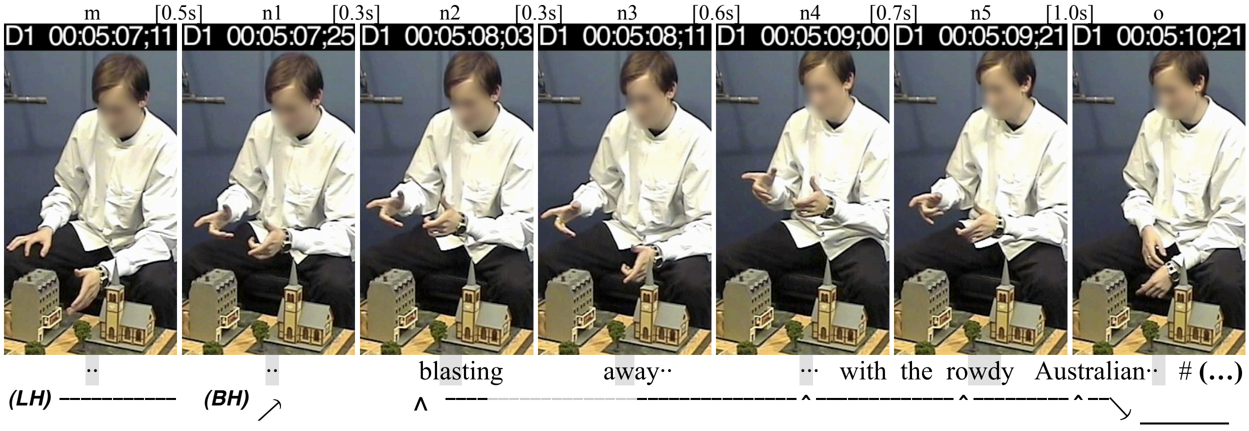
(3.5d) **A**: Okay ··· # ········· Yeah so··· Just to recap we'll* · #

I'll ·· be here with whoever I* ·· whoever I can get out of House 35,

m n1 n2 n3 n4 n5 o
·· ·· blasting away·· ··· with the rowdy Australian·· # hunter ·· soundtrack ··

m-n1: *RH has been on thigh while LH palm is holding in a deictic targeting House 35 (located just outside of frame), then BH rise in the same symmetrical configuration seen several times already, with lower three fingers of each hand curled while thumbs and index fingers are extended.*

n1-o *While keeping this shape, BH continue rising into a short stroke and then hold in place, except that LH sags significantly and almost drops back to thigh momentarily, after which the handshape becomes well defined again as BH rise into an upward beat during the pause (image* [n4]*), then are lowered for two downward beats on the stressed syllables of "rowdy" and "Australian." After the second beat BH immediately drop to rest for the remainder of the utterance.*


m [0.5s] n1 [0.3s] n2 [0.3s] n3 [0.6s] n4 [0.7s] n5 [1.0s] o
D1 00:05:07;11
D1 00:05:07;25
D1 00:05:08;03
D1 00:05:08;11
D1 00:05:09;00
D1 00:05:09;21
D1 00:05:10;21
·· ·· blasting away·· ··· with the rowdy Australian·· # (…)
(LH) (BH)

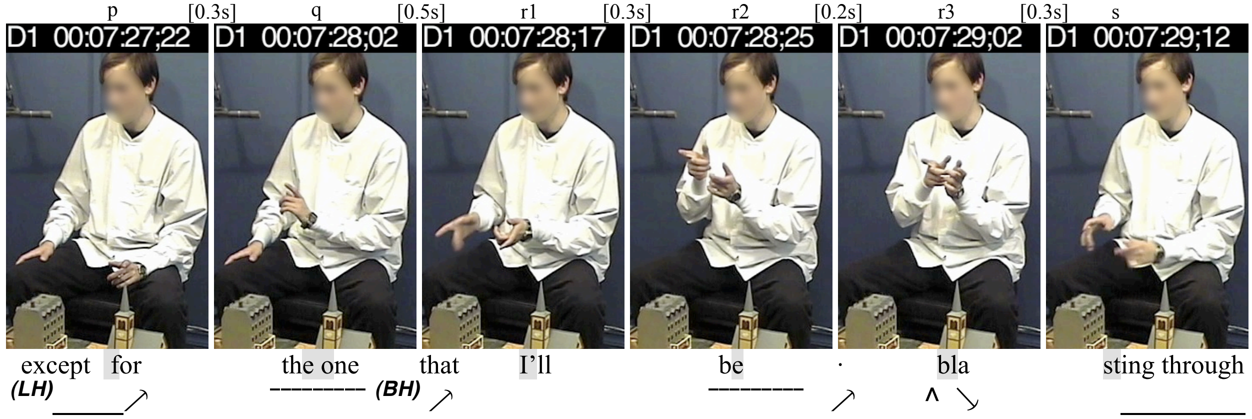
Although the preparation in 3.5d is very short, following the stroke he maintains the handshape for an extended period, with superimposed beats, as he pauses and then elaborates with additional associated speech. He drops the gesture in the middle of this elaboration, once again demonstrating that post-stroke holds do not have to encompass the entirety of the associated speech before dropping to rest.

Finally, 3.5e may be an illustration of the reduction in form that begins to take place as the catchment progresses further on the path toward becoming a label. It is worth examining whether this reduction in form also reduces the listener's ability to recognize the catchment. Alternatively, the change in how the catchment is gestured may actually flag it recognizably as a recurring form, leading to a stronger predictive inference by the listener regarding the speech that will go with it.

With speech leading up to it that has already been used in 3.5c (consisting of "except for the one that I'll…"), the final instance of the catchment is short in duration and begins dropping to rest almost immediately after the stroke, at the start of the word "blasting" (images [r3] to [s]). The hands are allowed to nearly come together in the center of the gesture space, rather than being held apart, and the stroke itself is a sort of sideways swipe rather than the clear jerk of the wrists that was used in previous instances. The repeat of the speech from 3.5c appears 1.5 seconds before the stroke and the word "blasting." The full gestural form of the catchment, meanwhile, becomes visible to the listener somewhat later (see image [r1]), but still a half second before the stroke and the appearance of "blasting" in his speech.

(3.5e) **A**: You tell me if they're coming out or not · and then · if they aren't, ·
I'll send in· whoever I have · from · House 35 in· to like · jus'·· scare them out. ····
B: Well we should at least ·· leave ··· guards · posted ·· at all places.
A: Well as long as I lock · the doors · to the theater
p q r1 r2 r3 s
except for the one that I'll be · blasting through · and then ·
# if I see any coming I'll just close the door.

p-q: *LH rises from rest into a vague index finger deictic pointed roughly toward his partner.*
q-r1: *RH rises to join LH as BH switch to the final instance of the catchment, which exhibits the three lower fingers of each hand taking on the familiar curled shape very early in the preparation.*
r1-r2: *BH continue rising and come close together; their movement slows nearly into a hold.*
r2-s: *After less than 0.2 seconds of nearly paused movement, BH are given a loose sideways stroke toward his left, immediately followed by a retraction that is complete before the end of "blasting."*



I am arguing for the expressive possibilities of the entire pre-stroke portions of some gestures, which, like holds, have generally been excluded from the set of phases considered to be 'expressive'. As discussed in Chapter 1, Kita et al. (1998) did explicitly include both pre-stroke and post-stroke holds as optional components of an "expressive phase" that includes the stroke, but the preparation phase was still not considered expressive in their model. It may be that the preparation phase is not generally expressive from the speaker's point of view, but a model exclusively focused on production misses out on the potential for this part of a gesture to reveal to a listener what a speaker is *about to* express explicitly. A speaker cannot withhold the informativeness of the body's configurations during the lead-up to the stroke, even if the stroke itself is withheld until being synchronized with particular elements of speech. Therefore, when the gesture's shape, and not just its stroke motion, is referentially significant, the gesture's expressiveness 'bleeds through' in spite of any intention by the speaker to withhold it.

As I argued was the case for 'prospective' gesture holds occurring near the start of disfluent utterances, the highly recognizable preparation to a catchment, once it is established through recurrence, could be a mechanism for an especially powerful kind of 'particularized conversational implicature' (Levinson 2000; see p. 49, footnote 10 in Chapter 2 above). They reveal not just a gestural form to go with impending speech, but the actual words that the speech will contain. A listener keeping track of a speaker's intended meanings will be able to make use of these cues to more efficiently interpret and respond to utterances. Catchments are unique to the discourses in which they occur, and their power is dependent on being situated following an evolution from repeated appearances. Via on-the-fly conventionalization with recurrent speech, they can be gestures that cue more than the shapes and relationships inherent in their forms, and any moment at which their form becomes recognizable is a moment at which the listener's process of comprehension could be strongly affected.

Partly because of this special mechanism of temporary conventionalization, through recurrence, of a speech-gesture *pairing,* I am motivated to keep the concept of *catchment* separate from the general phenomenon of gestures being extended via *holds.* The essential nature of catchments has been described as *recurrence* of gesture form ("recurrence is essential for a catchment," McNeill 2005, p. 219), which creates linkages across discourse and is one way in which coherence (or "cohesion") across time during communication is accomplished (this is the theme of the 5th chapter of McNeill 2005, on gesture and discourse). However, McNeill has also (personal communication) begun to consider gestures which *extend* (rather than recur) beyond their initial gesture phrases to be a form of catchment. This seems to differ from earlier statements in which extended gestures were described in a brief section as "layering with two hands," with gestures alternating between performing active depiction and being in "hold mode." (McNeill 2005, pp. 178-179).[10] McNeill presents this "layering" of extended holds as a closely related mechanism to catchments, at the head of his chapter (p. 164), because extended holds also create discourse cohesion across time in the gestural modality. However, this kind of layering is not there referred to as a form of catchment, and I will stick with the distinction.

[10] The alternation he discusses seems quite similar to the alternation in the chain of gestural 'buoys' I illustrated in Example 2.10 at the end of Chapter 2.

Now, I certainly agree with McNeill's emphasis that extended gestures create discourse cohesion, as I discussed at the end of Chapter 2, so perhaps these are just minor terminological issues.[11] However, using the term 'catchment' for holds as well as recurring gestures would seem to create a problematic unification of anaphoric *maintenance* with anaphoric *recurrence.* In the former, a significant cue can remain active in the gesture space by virtue of simply being there—it can be *self-evidential,* rather than *reconstructed,* as McCullough (2005, p. 633) notes. In the latter, the gesture is explicitly reenacted, and often synchronized with a repetition or near-repetition of the same linguistic items, which is not what occurs with extended holds.

McCullough (2005) discusses a phenomenon which is a blend of maintenance and recurrence: extended holds can be explicitly refocused multiple times throughout their lifespan, by way of superimposed beats, akin to the 'inertial beats' superimposed on a constant handshape discussed by Tuite (1993). Such refocusing beats can certainly indicate a recurrence of focus on a single gesture form, though it is a form whose scope is congruent with various different speech clauses, rather than speech which is itself recurring (as in the example I presented above). But extended holds frequently occur with no refocusing beats at all. We cannot be sure that all extended hold are being newly recruited as part of the production process of each subsequent clause; instead we can say that a hold is still there, and still consistent with ongoing speech, so that it can be unproblematically inherited from prior utterances.

Another reason to keep 'catchment' and 'hold' separate is simply the effective scope of their terminology. Within the term 'hold', we have stroke-dependent holds of the pre-stroke and post-stroke variety, 'independent holds' (Kita et al. 1998) not associated with an adjacent sharp movement—also referred to as 'hold-strokes' (Duncan 1996) or 'stroke holds' (McNeill 2005) to capture the idea that they serve the same role as strokes, and 'buoy' holds lasting across multiple clauses while other gestures are performed (Enfield 2004, 2009; cf. Liddell 2003). The simple notion of 'hold', then, logically contains all these types: it is simply the idea of a gesture form remaining the same across a span of time, regardless of length. A hold is *maintenance* of gesture form, generally speaking (McCullough 2005), and the plurifunctionality of these maintained forms, depending on how they line up with concurrent speech, has been the theme of this dissertation. The term 'catchment', in contrast, is much less self-explanatory. According to McNeill's (personal communication) most recent, most general characterization, it seems to capture the idea of gesture form being the same at two or more points in a discourse, regardless of what is occurring in between. This least restrictive use of the term would capture the generalization that maintenance and recurrence are *both* manifestations of discourse coherence across time. But because the most frequently discussed

[11] Except, as discussed in Chapter 2 (footnote 12 on p. 58), I find problematic the characterization of these layerings as a division of labor between a moving gesture's "depictive content" and a held gesture's "discourse content," because I do not agree that an extension of a gesture (e.g., in buoy form) immediately reduces its narrative-level, representational function. It becomes ground rather than figure, to be sure, and does indeed create cohesion across time, but this "discourse content" (if we wish to call it such) is in addition to, not instead of, the representational functions it performs in combination with other gestures. It may lose its representational functions when (and if) it remains in the gesture space long enough that the participants forget how and why it got there.

examples of catchments involve *re*-enactments of a gesture form (indeed *recurrence* has been stated to be "essential" to catchments), and these recurrences frequently occur in widely spaced fashion, I will avoid using 'catchment' to refer to extended holds, lest we lose sight of the potentially important differences between recurrence and maintenance.

Conclusion

In Chapters 2 and 3 I have given first a broad survey of 'prospective' and 'retrospective' gesture holds occurring with pauses and disfluencies in speech, followed by integration with several important studies describing the functional effects of disfluencies for speakers and listeners. Since disfluencies necessarily coincide with a speaker's commitment to utterance, while also leaving gaps in expression, the fact of having access to gesture provides listeners with a natural source of evidence for ongoing observation of a speaker's thought processes and intended meanings.

Disfluent utterances sometimes exhibit speech and gesture that appear strikingly asynchronous, especially when a minimal amount of speech is accompanied by a gesture and then followed by a hiatus, with the speech later elaborated more fully with content that matches that of the earlier gesture. However, in every case appearing to exhibit major asynchrony between the most obviously coreferential parts of speech and gesture, I have found that the elements which are actually co-timed are unproblematically co-expressive, with gesture and speech presenting complementary but readily associated content. The notion that every gesture must have a "lexical affiliate" rests on the false assumption that co-expressive elements of speech and gesture should express the "same" content, when in fact that is just one of the ways that co-timed elements can be semantically associated.

Nonetheless, speech-synchronized stroke phases are unavoidably preceded by preparatory phases, and under certain circumstances, such as when a specific gesture form recurs repeatedly with specific speech, preparatory phases can be extremely revealing of impending co-expressive speech and gesture. With such strong 'prospective' cues presaging the impending utterance, a listener may escape somewhat from the constraints of synchronous *interpretation,* even when speech and gesture *expression* are as synchronized as they can possibly be.

4 Gesture holds as 'bridges' across interruptions

Up to this point, I have focused on gestures spanning the natural disfluencies and pauses that occur whenever a speaker produces a stream of utterances. I now turn to disruptions of speech arising out of *interference* from outside forces, namely interjections by conversational partners, that cause speakers to suspend utterances in progress. It turns out that speakers' gestures often hold in place across these disruptions, just as they do across 'speaker-internal' disfluencies. I outlined some of the basic mechanical aspects of many gesture holds coinciding with disrupted speech, independent of the content of speech or gesture, in Chapter 2 (Example 2.1, pp. 37-39). My basic premise was that when a speaker intends to continue, an interruption in the flow of speech results in temporary silence but is often accompanied by something much more than gestural 'silence': the body frequently maintains a configuration across the hiatus.

In the case of the 'prospective' gestural cues discussed in Chapters 2 and 3, the gestures in question were coreferential with, and lasted long enough to eventually coincide with, not-yet-realized speech (meanwhile they were also co-expressive, in complementary fashion, with the most immediately co-timed speech). They were shown to be capable of helping a listener predict the most likely intended content, sometimes inducing the listener to offer a sentence completion during the speaker's hiatus, at least when coupled with listener-directed gaze (Examples 2.3 and 2.4, pp. 43-49). Prospective gestural cues can also render a speaker's disfluent utterance sufficient, from the point of view of the listener, even when the aborted speech stream is never repaired (Example 2.2, pp. 40-43). Meanwhile, in the case of the 'retrospective' gesture holds discussed in Chapter 2, the gestures coincided with fully realized speech, but continued to contextualize ongoing utterances in the form of holds lasting across multiple clauses and through pauses and digressions in speech.

Throughout those examples, I focused primarily on the experience of a listener trying to make sense of a speaker's message. Lacking access to a speaker's internal utterance formulation processes, a listener is left to generate, as quickly as possible based on available signals, a constantly evolving guess as to the speaker's intended message.[1]

[1] As evidence for the listener's rapid process of fixating on the most likely referents based on verbal and non-verbal cues, in Chapter 3 I cited several eye-tracking studies from the psycholinguistics literature whose results seem readily extendible to the conversational examples from my corpus (see Chambers et al. 2002; Arnold et al. 2003, 2004).

However, the speaker's experience of his or her own maintained gestures could be equally important, particularly when an utterance is disrupted by outside forces. When the attention of both speakers *and* listeners is drawn forcibly to new stimuli while an utterance is already in progress, it could be particularly advantageous for part of the speaker's body to be left in a configuration tying it to the recently disrupted and still incomplete utterance. This would still apply when it is the listener who is responsible for the interruption—it would be advantageous, at least from the point of view of the original speaker, to have a way of automatically reorienting all participants back to the material being discussed previously. In this way, gesture holds spanning interruptions could serve as short-lived 'cognitive artifacts' (Norman 1991; Hutchins 1996; Enfield 2005, 2009), operating within the brief timescale of a few utterances. While not usually spanning more than a few seconds of discourse, a gesture hold surviving beyond a short but disruptive period of interference could still provide a crucial link to prior discourse. Depending on the referential power of the gesture's form—for example, if its shape or deictic target are strongly associated with entities, referents, or topics under discussion before the disruption—a remaining gestural artifact could automatically and unavoidably contextualize the resumption with the same referential power.

Crucially, this mechanism would not require continuity of attention, or continuous activation in working memory, precisely because of the use of the body as artifact: relevant features can survive in a 'material carrier' (McNeill 2005, from Vygotsky 1986) even if all of one's attentional cognitive faculties are temporarily occupied with other matters. This is what inspires my particular interest in gesture holds as 'bridges' across interruptions. Later in this chapter I will lay out in greater detail how gesture holds (whether performed or observed) may affect attention and memory during speech.

Interruption-bridging holds in commonality with other holds

A case of an interruption-bridging hold was already presented in Example 2.8 (pp. 54-57). In this example the speaker maintains a deictic hold on a house in the model village ("We have to go this house?"), and then makes a metanarrative aside ("Yeah, something like that"), itself a fluent form of self-interruption. As she is about to continue, she refocuses her deictic gesture, but then she and her interlocutor are unexpectedly caught up in laughter. After this interval, she continues where she left off, with her gesture conveniently still indicating the referent she had been in the process of discussing. There can be many kinds of interruptions from many sources, such as distracting stimuli from the environment (street noises, natural phenomena, loudspeaker announcements, etc.), unexpected emotional outbursts such as the laughter that frequently beset the participants in the "wombats" study, or competitive interjections by others. In the current chapter, my focus is on the last of these: gesture behavior during periods of interference or distraction that occur because of competitive or collaborative speech by conversational partners.

Besides providing an artifactual cue linked to prior content, a gesture hold lasting across a competitive interjection will also show the original speaker's intention to resume speaking (see Schegloff 1984), just as when a hold is maintained across a silent pause in speech. There are a great many commonalities in form between hiatus-spanning holds occurring with pauses and disfluencies, and the ‘interruption-bridging’ holds I discuss in this chapter. In both sets of cases, the hold shows an intention to continue speaking, and also acts as the basis for continuity by maintaining a configuration that will relevantly contextualize the resumption. The speaker’s eye-gaze is also usually directed at the gesture, or off to the side, rather than at the interlocutor, which accentuates the speaker’s *lack* of desire to elicit a response from the interlocutor (see Kendon 1967; Stivers and Rossano 2010). But interruption-bridging holds are also rather different in a few respects from personal hiatus-spanning holds: an interlocutor attempting to take over a speaker’s turn has often had enough of the speaker’s intended message, and wishes to add a new independent contribution. From the original speaker’s point of view, dealing with a partner’s interjection could also be more disruptive to utterance production processes than an ‘internal’ instance of hiatus, because it will require attention to be directed toward the other’s utterance. In contrast, many personal disfluencies in speech are of a purely ‘surface’ variety (see Kita 1993) and do not involve competing messages. For example, as discussed in Chapter 3, Clark and Wasow (1998) claim that pausing after the start of an utterance can be a normal step in utterance production, allowing a speaker to commit to speaking while also providing time to fully craft the impending message. In contrast, unpredictable interjections by interlocutors could completely disrupt a speaker’s plans.

For these reasons, in considering interruption-bridging holds, it is especially important to consider the experience of all parties involved rather than focusing on just the original listener or speaker. Indeed, such holds span periods when *both* participants are attempting to play the role of speaker. A person maintaining a hold must contend with the distraction posed by the competitive utterance, and must also assume that the wayward, interrupting listener is switching attention away from the original utterance that was in progress. The externalized artifact of the hold, performed by one person but accessible to both, could therefore be a useful cue for getting both participants ‘back on track’, from the point of view of a speaker wishing to resume a disrupted utterance.

From a slightly different perspective, interjections by the listener can often be considered collaborative rather than disruptive, to the point that the progress and authorship of utterances become extremely complex, in terms of the overlapping speech and gesture of both participants. For the same reasons already discussed, gesture holds by one participant or another can become collective tools of coherence, as illustrated in Example 4.4 later in this chapter (pp. 114-116). Meanwhile, Example 4.6 at the end of the chapter exhibits intertwining collaborative and competitive elements (pp. 119-122).

Example 4.1 — a ‘prospective’ deictic surviving across interruptions and restarts

This first example shares certain features with Example 1.1 from Chapter 1. In both cases, the moment in question involves a speaker aborting and restarting some version of the phrase “we should” while maintaining a deictic gesture hold. But in the case below, a complicating element is that the first hiatus and restart in Participant A’s speech is caused by his partner’s interjection. The moment in question is shown in the images and transcript from [c2] to [c4], and discussed in further detail following the example.

(4.1) **B**: We could · lay some charges like · up on the roof area here? ···

⌈**A**: But we don't wanna* ··· ⌉
⌊**B**: And then while* · like ·· I'll · lay some ch* ⌋

A: We don't wanna blow up the whole movie theater.

B: I know! But you're talking about using explosives. {*sigh*}

A: I'm talking about small explosives that don't really blow up that much but are just really loud. ····· And we just ··· put some* ····

⌈**A**: You know like those fireworks? ⌉
⌊**B**: Well if we're trying to drive'em ··· [a]out here, ·· # ⌋

B: We need to put'em up her[b]e. ··

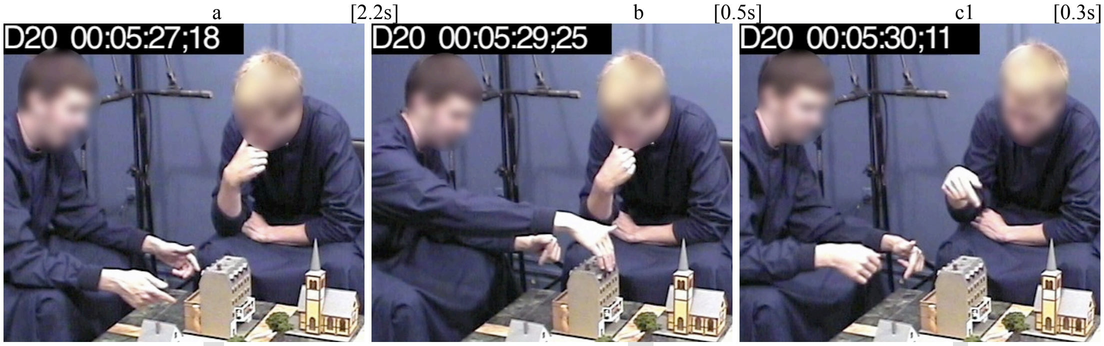
⌈**A**: Yeah wel[c1]l what I'm s[c2]aying is we should* [c3]· # we should do it* ⌉
⌊**B**: In front. ⌋

A: we should make it [c4]a line thr[d]ough the rooms.
So that they explode like here then here then here then here

⌈**A**: and like drive'em out to the front.* · Out the back. ··· ⌉
⌊**B**: Yeah but if you're talking about* ··· ⌋

B: Alright. ···· Well that's fine with me.

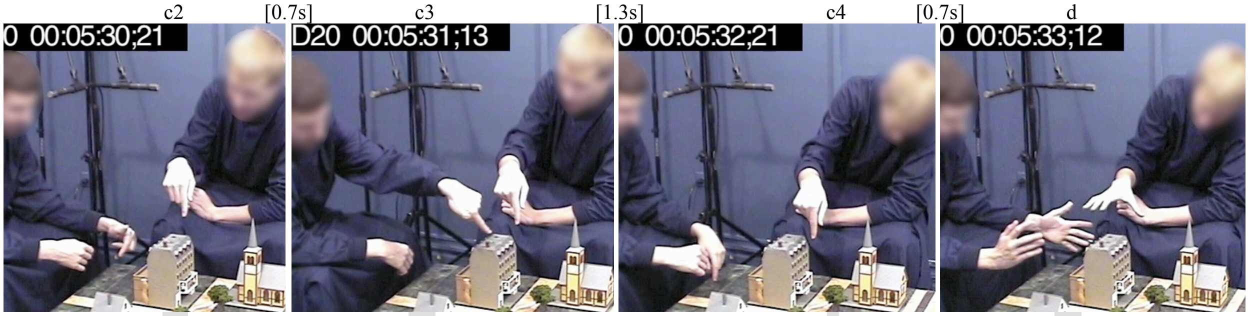
a: *Participant B has reached out with BH to rear of theater, and flicks fingers toward himself twice.*
a-b: *While leaving LH held in place, he reaches out RH to sweep in a line across top front of theater.*
b-c1: *As Participant B retracts RH back toward rest, Participant A begins speaking and starts to extend his RH toward theater, with index finger pointing down.*



B: if we're trying to drive'em ··· out here, ·· # We need to put'em up here. ·· |**A**: Yeah well

(A's RH) ___↗
(B's RH) ___↗ ————^—^————— ↗ ^ ↘
(B's LH) ___↗ ————^—^——————————————————————————————

c1-c2: *Participant A's RH become held in place with index finger pointing down at top front of theater.*
c2-c3: *While Participant A's RH remains held in place, Participant B uses LH index finger in a similar shape to sweep in a line at the top front of the theater where Participant A is already pointing. Participant A stops speaking momentarily just after his partner's briefly competitive speech.*
c3-d: *Participant A maintains his hold through two resumptions of speech, then spreads his fingers into a dome over entire theater (and Participant B opens BH briefly as well, at almost the same time).*
c4-d *Participant A tilts his head down and to the side to peer at the windows at the front of the theater.*



⌈**A**: what I'm saying is we should* · # we should do it* we should make it a line through the rooms. ⌉
⌊**B**: In front. ⌋

(A's RH) ———————————————————— ↗ ∧ ————————
(B's RH) ____________________________________ ↗ ∧ ———— ↘ ___
(B's LH) —————— ↗ ∧ ——— ↘ ______________ ↗ ∧ ———— ↘ ___

At image [c1], Participant A (seated on the right) begins a deictic gesture formed by his right-hand index finger pointing straight down at the top-front area of the movie theater. Following the deployment of this hold, during the two restarts in his speech, the form of his gesture remains constant: it is simply maintained. But the nature of the two suspensions and restarts is quite different. The first suspension is a reaction to his partner's interjected speech and gesture, which are themselves an addendum to an earlier utterance. The interjection consists of the words "in front" coupled with a brief deictic gesture pointing to almost the same location as Participant A's ongoing hold (the striking similarity of their gestures may illustrate a case of 'mirroring', but it could also be that they are simply both performing a canonical index finger deictic). Just after the interruption, at the moment of image [c3], Participant A stops speaking, pauses and inhales, and only resumes when it is clear he is no longer being interrupted. His second restart (between images [c3] and [c4]) appears to be a reaction to his own decision to change his wording, rather than to any direct action by his partner, although his disfluent delivery could perhaps come as a result of being discombobulated by the earlier interference.

In Example 1.1 from Chapter 1, the speaker's deictic was held through his disfluent repetitions ("we sh we'll we'll…") at the start of an utterance. Similarly, the speaker in Example 4.1 breaks off his speech just after saying "we should" while gesturing to a location, and he has likewise not yet produced a verb phrase following "we should." This time, however, the listener is competing with his utterance, and is contributing to the established topic (the question of the type of explosives to use, and how they should be deployed). By maintaining his held gesture through his partner's contribution, Participant A signals his *prospective* intent to resume speaking, meanwhile his aborted speech simultaneously indicates to his partner that he has paused and is *currently* listening. In addition, the recent linking expressions in speech ("yeah well what I'm saying is…") indicate that the future resumption will be related *retrospectively* to the content they have both been speaking about most recently. Now, I am certainly not claiming that Participant A would have been in danger of losing track of his topic in this instance, had he been without the aid of his gesture hold during his partner's interruption. If there is a universal mechanism causing the body to maintain configuration across disruptions, we should expect it to apply across a broad range of disruptiveness, including cases where some of its benefits are unnecessary. When a speaker intends to continue, a gesture hold in any case of interruption is completely consistent with resumption—there is no possible disadvantage, meanwhile there are the various possible advantages I have discussed, none of which needs to be operative in every instance.

Example 4.2 — a 'retrospective' iconic with partial retraction and 'prospective' resumption

In this example, the participants have recently been discussing the plans which came into focus above, in Example 4.1, which will require them to enter the front of the movie theater and drive the wombats out the back with the aid of noisy explosives. At the start of the transcript below, they then begin joking about recruiting more extravagant helpers, such as a full SWAT team and even the Crocodile Hunter himself. Participant B then attempts to return to the 'serious' discussion they had been having before, and says "I mean I think, I don't know I think that's a good idea…." When he says "I think that's a good idea," he is referring to their 'serious' plan, but his partner takes it as an opportunity to interject, jokingly, that he is referring to the Crocodile Hunter instead. The images illustrate what happens to Participant B's gestures as he navigates the interference.

Just before the interruption, he creates an iconic representation which associates straightforwardly with "get'em surrounded" in his speech (see image [b1] below). His slowly decaying gesture, 'retrospective' to this clear association, is then partially maintained as he contends with his partner's interruption and his own responses to it. During his partner's interjection (consisting of "What, the Crocodile Hunter?"), Participant B aborts his speech but continues to maintain the 'surrounded' shape with both hands. During the silence that follows, his muscles begin to relax (see images [b2] and [b3]), until he lets his wrists go completely limp while jokingly agreeing with his partner's suggestion (see images [b3] to [c]). After less than a second of limp muscle tension, however, he reinitiates the same 'surrounded' configuration as before, while refuting his own joking assent, letting out a laughing exhalation, and inhaling to resume speech (see images [c] to [d2]), all of which are clear *prospective* signs of a resumption. The resumed handshape itself, meanwhile, unmistakably links the resumption *retrospectively* to his activity just before he was derailed by his partner's joke. Just as I argued for Example 2.8 in Chapter 2, which I briefly reintroduced at the beginning of the current section, in Example 4.2 Participant B's gestures are a combination of attempted maintenance across a digression, followed by strong refocusing ahead of a resumption. This time, however, the digression is caused by his partner's interference, rather than a decision by the speaker to insert an aside.

(4.2) **A**: I think what we need is a SWAT team too.

B: Yesss! |**A**: @@@ |**B**: Yessss. |**A**: No. |**B**: No, for real. ···

B: ’Ey, we need the Crocodile Hunter · is what we really need. |**A**: @@@

B: I mean, no · seriously · I think we need* ·· # I mean* ··

A: The Wombat Hunter. ···· |**B**: The Wombat Hunter. ····

⌈**A**: Well* ⌉
⌊**B**: I mean I think* · I don’t know I think that’s a good idea I think it will work and uh, #⌋

a b1 b2
⌈**A**: What, the Crocodile Hunter? ·· ⌉
⌊**B**: That should get’em out here and we get’em · surrounded and* ⌋

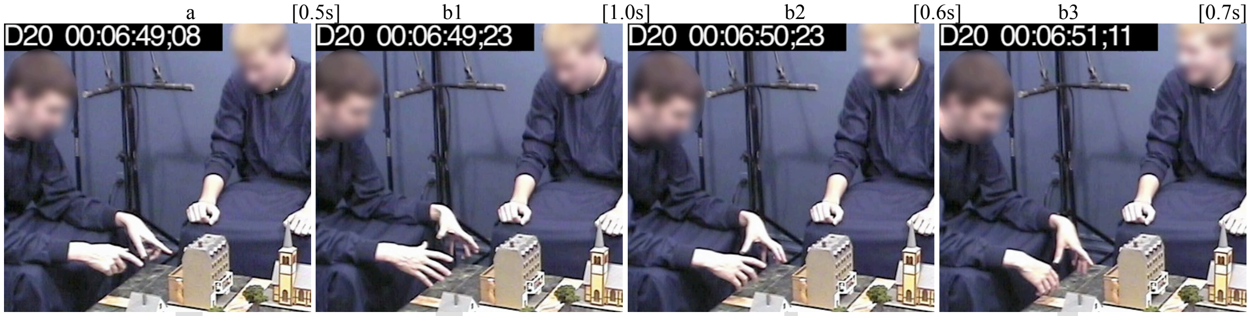
b3 c d1
B: # Yeah. ·· |**A**: @ |**B**: Nah @ #

d2 e
B: We get’em right here we get’em surrounded and ··· *(snaps finger)* ··· take’em out. ··

B: Taser’em a little bit. ··· |**A**: Tase’em. ···· |**B**: Maybe · eat one or two? ·····

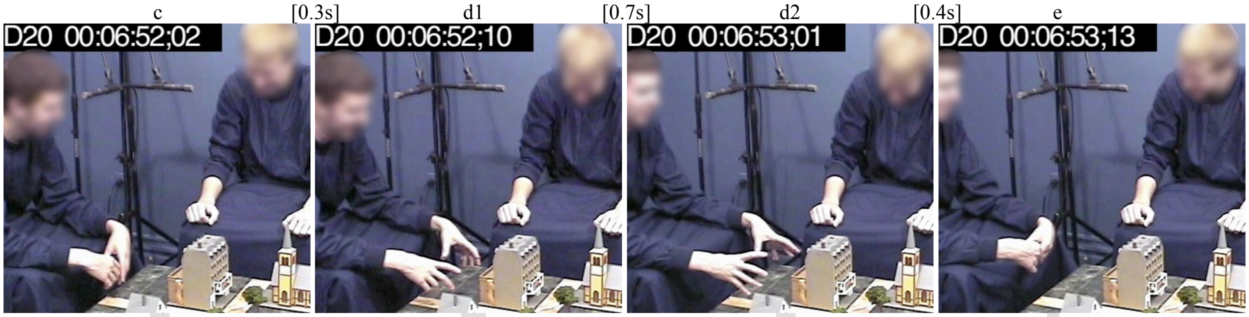
⌈**A**: Are you some kinda savage? ⌉
⌊**B**: Heard wombats are really tasty. @ @ ⌋

a: *Participant B has had hands at rest in front of knees, but in the lead-up to image* [a] *he has weakly extended RH index finger twice, on “that” and the first “get’em” (not shown). He then extends BH index fingers toward rear of theater and holds the configuration.*
a-b1: *BH then spread fingers into a dual-handed dome shape over the area behind the theater.*
b1-b2: *This config. is maintained as Participant A interjects speech, and Participant B suspends speech.*
b2-b3: *During the silence, Participant B’s hold eventually drops part way toward rest at the wrists, but as he inhales before jokingly responding, his LH index finger regains some of the dome shape.*



⌈**A**: What, the Crocodile Hunter? ·· @ ⌉
⌊**B**: **(...)** and we get’em · surrounded and* # Yeah. ·· ⌋
(B’s BH) ↗ ———————— ↗ ———————————— - - - ——— ↘ ______

b3-c: *BH wrists and fingers eventually become fully retracted (for only about 0.1-0.2 seconds).*
c-d1: *As he gives cues of a resumption of 'serious' speech ("nah" followed by a laughing exhalation), BH regain the same dome shape as before.*
d1-d2: *His fingers become spread further into a more well-defined dome shape, as the hold continues.*
d2-e: *BH are retracted and clasped together. (After image* [e], *the fingers extend again momentarily, and the index fingers sweep down in a stroke on "surrounded" – not shown.)*



B: Nah @ # We get'em right here we get'em **(...)**
__________ ↗ ––––––––––––––––––––––––––––– ↗ –––––––––––– ↘ __ ↗ ––––––

Unlike many examples I have presented, this one involves a noticeable *recurrence* of a recently performed gesture, rather than simply maintenance, because Participant B nearly disengages his gesture as he momentarily humors his partner's joke. It is therefore akin to the 'catchment' structures explored by McNeill (2005, ch. 5), a topic I discussed at the end of Chapter 3. Once the thread to his discourse before the interruption is reestablished, he drops the handshape and clasps his hands for a moment (image [e]), then reiterates the shape again, in reduced form on "surrounded" (not shown, but see the full transcript at the start of the example), before finishing with finality via a right-handed finger snap.

In another case which I will not illustrate with a full example, this same speaker begins to gesturally 'bridge' an interruption from his more dominant partner, and holds his configuration for a short time, but then abandons the attempt and allows his partner full control. The decision to withdraw the hold appears to coincide with his choice not to resume his prior utterance: he abandons his "claim to 'speakership'" (Schegloff 1984, p. 272). While I have suggested that a speaker can enter into a hold partly as an automatic response to interruption or disfluency, the maintenance of a hold also broadcasts an intent to continue, and is *inconsistent* with ceding the floor to new topics with different referents: recall Kita's (1993) findings that gestures were very unlikely to survive across utterance repairs that altered or invalidated the image created by the gesture. If there is an automatic 'hold response' to interruption or disfluency, it can quickly be canceled if maintenance of the hold becomes inconsistent with the speaker's evolving role or plans.

Example 4.3 — a 'buoy' lasting across multiple contributions by both speakers

The transcript for the next example is extremely long, so I have enclosed in boxes the discontinuous portions which are lined up with the images. In contrast, the gesture hold I wish to illustrate is actually very simple: it is a buoy lasting ninety-seven seconds.

(4.3) **A**: So, ···· so, ····· here's the church, · and here's the movie theater · right? ··· the church ········ is actually*
So, · the church is here · and the movie theater is actually kind of lo··ng··

[**A**: Right here. The entrance is actually··]
[**B**: Long. · Which part is the entrance? ····· on the street? ···]

A: It's not on the* · It's not on the street, the doors · are actually · on the other side.

[**A**: It's not · the one facing the street. The* Right. It's* It's that]
[**B**: On the other side. Oh it's not the one facing the house either.]

A: Exactly. ·· 'cause there's a bunch of windows there. |**B**: Okay.

A: And when I looked at* · when I looked at the plan, ····· the doors · were actually on the other side.

[**A**: Right. · Exactly.]
[**B**: On* ·· As if they were coming in from the back. ·· From the back of the street. · Okay.]

A: And then there's like a little terrace? |**B**: Uh-huh. ·· |**A**: Right above the doors? ·· And there's

[**A**: like a little terrace, ····· on the other side too so on both sides there's like a mini terrace.]
[**B**: A matching terrace.]

[**A**: # Right. A matching terrace. And I think* ··]
[**B**: Okay. And the windows are the sides that don't have a door? ····]

A: The windows are on both sides. ·· |**B**: That don't have doors or do have doors? ··
A: They* · The windows are on s* on the side that do have doors, · but about half the amount.
Not eighteen so about nine or ten. ····
And ·· there's actually eighteen windows on the other side that faces the pile? ········

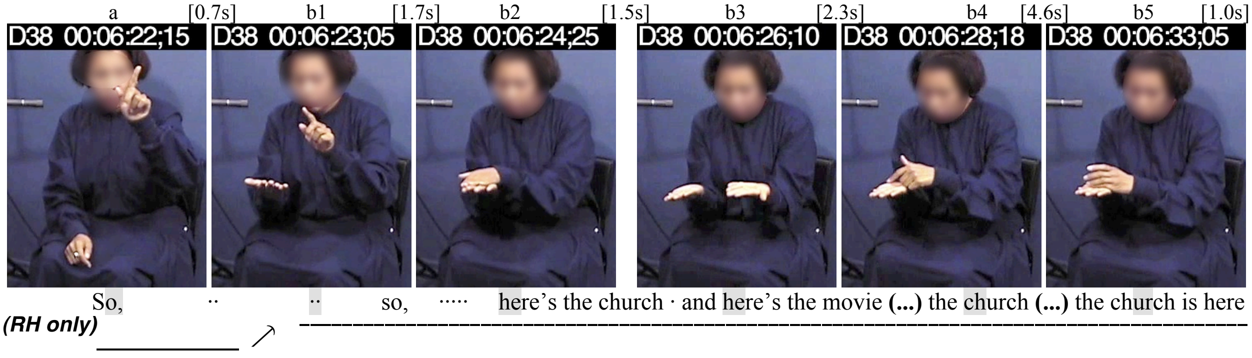
[**A**: Do you see what I'm saying?]
[**B**: Okay, so, the ·· side · that has the terrace without the door faces the pile? ···]

A: The side that has the terrace without the* Yess. |**B**: Okay. ······

[**A**: Yes. · And the side* @@ Exactly. # @ Wow, @ # # I don't know@ here, # @@ #]
[**B**: Okay @ @@ # @@@ #]

[**A**: So anyway*@ ·· # Uhm, so let's see ·· So what we need to do, with all of this information,]
[**B**: Okay so* Uh-huh?]

a-b1: *RH rises into a flat palm with unspread fingers, palm facing straight up.*
b1-b5: *While RH remains held in place, LH palm pancakes onto it to indicate church, then flips over next to it to indicate theater. LH index finger points to RH church buoy, then whole LH places it again.*

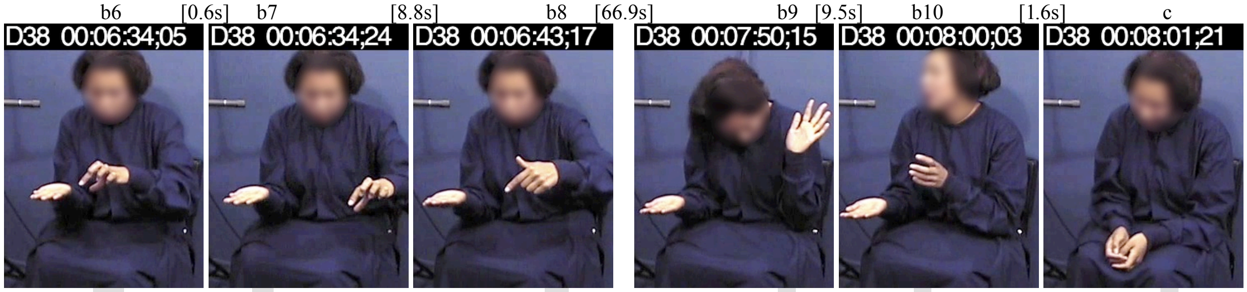


b5-b7: *LH then returns to gesturing about the theater, and no further remarks are ever made about the church. LH thumb and index finger track a long building outline adjacent to the church buoy (images* [b6] and [b7]*), and later LH is used to pinpoint theater's entrances, as in image* [b8].

b8-b9: *For more than a minute of continued discussion with her partner (see transcript), LH continues to be used to gesture about sundry details of the theater, such as windows and terraces, while RH remains held in place as a buoy for the adjacent, unmentioned church. She is eventually overcome with embarrassed laughter after her partner's neverending insistence on additional details (image* [b9]*), but the RH church buoy remains stable.*

b9-b10: *She makes a major discourse shift ("So anyway," at* [07:54;00]*), but continues holding the RH church buoy for an additional six seconds even after the major topic shift in speech.*

b10-c: *Finally, BH are swept up and to her sides in symmetrical arcs which she lets fall heavily against her thighs (not shown), after which she draws her arms together to clasp BH at rest in her lap.*



and the movie theater is actually kind of lo··ng **(...)** the doors · are **(...)** Wow, @# **(...)** with all of this information,

––– ↗ ^ ↘ _____

Throughout this lengthy exchange, Participant A maintains a right-handed gesture buoy consisting of her flat palm facing the ceiling, which she established early on as the location of the church, by way of a strong pancaked clap from her left hand (see image [b2]). For more than a minute and a half, her 'church' hand remains completely immobile, as she and her partner discuss various details of the adjacent movie theater.[2] There are numerous disfluencies and a great many interruptions, especially by Participant B. Following the moment shown in image [b5], they never discuss the church at all; her gesture buoy is simply a rock-steady landmark adjacent to her left-handed gestures referring to the theater. Her partner's thirst for details is so insatiable that Participant A eventually hits an emotional peak (see image [b9]), after which she is finally able to move the discourse back to the issue of actually extricating the wombats. But for several seconds, during the lead-up to the moment of image [b10], the buoy is maintained even though she has shifted away from discussion of architectural details and back to their instructed task. The hold is not inconsistent with that discussion (which is about events which will take place in the town), but it is not needed either, and she eventually drops it during a return to symmetrical gesturing (just before image [c]). The buoy's extended presence in prior discourse may have rendered it almost completely forgotten and nearly

[2] Note that the model is not present for their conversation, so Participant A must recall everything from memory. She has either placed the theater inaccurately in relation to the church, or she has switched to an unusual perspective for the sake of convenience (her placements here are only accurate if her point of view is the same as that of the camera in most of the examples in these chapters, which is from the opposite side compared to how she observed the model). Regardless, her accuracy does not affect my discussion.

glued in place, though the speaker may have retained it in part as a (largely futile) attempt at broadcasting a wish for greater control of the discourse.

In this example I have not aimed to illustrate the microtiming of a gesture hold compared to concurrent speech. Instead, the example shows a hold deployed as an ongoing resource which pegs the gesture space as a mapping space for the absent model village, and remains stable across many varied exchanges between the speakers. Given the trajectory of their discourse, the importance of the hold lies not just in the stable positioning of the church building, but in the consequently stable position of the empty space adjacent to the buoy gesture, which becomes an invisible canvas upon which various features of the movie theater are depicted. For example, images [b6] to [b7] show the speaker outlining the length of the theater with her thumb and forefinger, and image [b8] shows her reaching over the invisible structure to pinpoint its doors on the far side (see the full transcript at the start of the example for the relevant context).

Theoretical Integrations of gesture with attention and memory

A paragraph at the end of Kendon's (1972) early seminal paper hints at two important possibilities for the relationship between gesture and memory. I am here referring to the "working memory" that speakers must keep active and accessible for the purposes of tracking what they are in the process of speaking about (Kendon cites Miller et al. 1960; see also Baddeley 1986, 2001; Cowan 1999; Baddeley and Logie 1999). This includes complex relationships between various entities, events, and contexts in different modalities such as the visuospatial and the verbal—that is, the substance of thoughts which must be 'kept in mind' and manipulated against each other as we formulate and comprehend utterances. One possibility is that this working memory, necessary "for the execution of speaking plans," could become partly "represented in" the postures and sequences of body positions taken on by speakers (Kendon 1972, p. 206). Kendon also suggests a potentially separate possibility, which is that "these body positions and the different movement sequences perhaps can function as a means of *storing information* about which stage of the plan is in operation" (p. 207, emphasis mine). Although Kendon probably did not intend for these to be separate perspectives, they can be interpreted in two quite different ways. On one line of reasoning, gesturing may help call up information stored in working memory, or keep it activated: gestures are seen as representing, facilitating, or strengthening a cognitive process. But on the other line of reasoning, gestures could be helping by *taking the place of* a cognitive process: instead of keeping certain memory traces active, they make it unnecessary for effort to be expended in the service of maintaining certain information in working memory, because the information is being stored in material form. To the extent that gestures or other material arrangements are able to last as artifacts, they can store a representation and keep it available across time to visual perception (and in the case of a speaker's own gestures, to kinesthetic perception).

The first perspective, in which the contents of working memory become represented in gesture, is taken by authors such as De Ruiter (1998), who argued that iconic gestures are "generated from imagistic representations in working memory" (p. 22), or are “derived from spatio-dynamic working memory” (p. 26). De Ruiter showed that when presented with a set of shapes and lines on a screen, and given the task of describing their arrangement, speakers gestured significantly more when speaking from memory than when the arrangement was kept available to them on the screen. These findings lent support, De Ruiter concluded, to the hypothesis that “gesture helps in retrieving spatial information from memory” (p. 55), and that facilitating such access was therefore one of the functions of gesture. In a related but independent study, Wesp et al. (2001) similarly found that participants gestured more often when describing a painting from memory, and they concluded that gestural movements help to sustain spatial representations in a working memory buffer (see discussion by Morsella and Krauss 2004).

The results, I believe, are also consistent with a different hypothesis. When De Ruiter’s (1998) on-screen spatial arrangement or Wesp et al.’s (2001) painting were available during speakers’ descriptions, their details could be referred to at any time, regardless of the state of activation of the participants’ spatial working memory buffers. The advantage of artifacts is that they can store information on their own, for example as “physical, perseverating embodiments of discourse entities and themes” (Duncan 2008, p. 306, in reference to gestural artifacts). They can take the place of some of what would go into a memory buffer, and their continued availability does not require effort or attention: information can be accessed sporadically without danger of being ‘forgotten’, because it is stored in a way that is immune to the decay of memory. In De Ruiter’s (1998) and Wesp et al.’s (2001) studies, the fact that participants gestured less when the artifact was available suggests that there was less need for gesturing in that condition; some of the benefit that gestures could bring to the task was already provided by the diagram or painting. This last point would not be disputed by the authors. However, instead of the idea that gestures mainly helped facilitate access to (De Ruiter 1998), or keep activated (Wesp et al. 2001) working memory traces of the details of the pictures, it might be the case that the gestures *recreated* and kept available*,* in quasi-artifactual form, some of the important structures visible in the pictures. Even if part of the effect of gesturing may be to initially reactivate a spatial memory, any information contained in the arrangement of gestures or other artifacts *across time* allows for some of those memories to be neglected rather than strenuously kept active. This would allow attention to be directed instead at material that is not being represented by external artifacts.

Another possibility is that a combination of these perspectives is correct: gesturing could help facilitate access to spatial memory by schematically standing in for some of the previously observed arrangements (see the section on *recall cues,* below). Once accessed, some of the gross structure of those arrangements could then be preserved across time, artifactually, in gesture, allowing attention to be devoted to consideration of additional details which depend on the basic structure being kept available.

The key here is our restricted attention capacity: when voluntarily directed, it is a unitary resource with a singular, rapidly switching focus (Pashler 1997). Any focalization of attention, regardless of target and regardless of modality, means a *lack* of focalization onto other alternatives. Furthermore, attention is inextricably bound up with working memory, because in the absence of other means of storage or activation, working memory degrades rapidly and cannot be sustained without *internal rehearsal* (Cowan 1988, 1999; Baddeley 2001), and internal rehearsal is itself a task which cannot be accomplished without attention being devoted to it. Given the large amount of information we must synthesize during communication, there is a very clear motivation for us to find ways of keeping some information accessible *without* the need for internal rehearsal, as well as to develop techniques that facilitate the retrieval of previous perceptions. During a task, anything allowing for less extensive demands of attention onto parts of the task will allow attention to instead be devoted more frequently to other aspects of it. Additional attention can be paid to rehearsing other details, or formulating future utterances, or attending to an interlocutor's contributions. The result is improved performance overall.

There is evidence that gesture allows for exactly this sort of alleviation of task-based demands on attention. Goldin-Meadow et al. (2001) discovered that speakers who were tasked with explaining a math problem at a blackboard, while also remembering a list of random letters, performed significantly better at the memorization task if they gestured while explaining the math problem. An illustration provided by Goldin-Meadow (2003, p. 152) shows the kind of gesturing involved: speakers pointed to the different sections of a pre-written equation while discussing their reasoning, a step-by-step process whereby the stages of their mathematical process seem to map fairly closely to the limb movements marking out each part of the equation, while the hand and fingers indicate linkages between the numbers and variables. Goldin-Meadow and her colleagues stated that "gesturing appeared to save the speakers' cognitive resources on the explanation task, permitting the speakers to allocate more resources to the memory task" (Goldin-Meadow et al. 2001, p. 516), but they did not discuss what those resources might be. They opted instead for the non-specific assessment that gesture "lightens cognitive load."

I would like to suggest a more specific explanation. To achieve better results on the memorization task, the participants must have been able to conduct internal rehearsal more effectively. Their environment, their speech, and their gestures were all related to the math explanation task, and completely disconnected from the memorization task. Yet gesturing during the explanation evidently allowed their attention to be devoted more frequently to internal rehearsal of the random letters. In other words, because of their gestures, the explanation task required less frequent focusings of attention: the structuring effect of their gestures appears to have 'offloaded' some of their cognitive demands into the material domain. The most straightforward way for this to occur is for some of the structure of their task—for example, the stages in the explanation relating to each section of the equation on the blackboard—to become effectively stored in their body positions, just as Kendon (1972) suggested. With each stage of the explanation linked by their bodies to the permanent arrangement on the blackboard, some of the structure of their "speaking plans" would be rendered immune to the decaying effects of time. Having each gesture serve as an anchor through the duration of each step, they could then devote attention more frequently to rehearsing the unrelated list of letters.[3]

Such gestural anchors, under direct control by the participants and thus much more agile and rapidly changeable than the material world around them, may still provide far greater stability across brief intervals of time than anything achievable by thought alone.

[3] Note that I am not claiming that gesturing frees up working memory "capacity" (cf. Duncan 2008); instead I believe artifactual gestures free up the ability to use more frequent attention *for* working memory, to more successfully keep any of its contents active, regardless of origin or modality.

Their benefits are not necessarily gained by way of sustaining material in a buffer of working memory, as Wesp et al. (2001) believe; continuous maintenance of working memory traces should not actually be required, when part of the supposed contents of these buffers is sustained materially instead and can be accessed whenever needed, or whenever a speaker's attention is not devoted to other more urgent targets. A process that offloads part of a task into the material realm can be thought of as a kind of automatization, allowing attention to be switched away to other matters without interfering with the automatized task. As discussed by Peters (1990), this may be an essential element of "doing two things at the same time," given the restrictively unitary nature of attention.

This perspective, in which certain cognitive tasks can be offloaded into physical means of information storage such as gesture, is also an integral part of the theory of 'distributed cognition' as developed by Hutchins (1995, 2006). Key to understanding this perspective is that information-bearing structures are distributed not just across the environment and bodies of the interacting parties, and the tasks they find themselves engaged in, but across time as well. On a long-range temporal scale, culture and cultural institutions create inescapable continuities between the past and the present; embedded within these, on a shorter-range temporal scale such as that within and across utterances, temporary artifacts create short-term continuities as well. Gestural artifacts act not just as images themselves, but as deictic tools highlighting particular elements of the environment and recruiting them for cognitive work by multiple individuals, as in Hutchins' (1995, 2006) analyses of the "environmentally coupled" gestures integral to navigation tasks on a Navy vessel (and cf. Goodwin 2007). These perspectives on gesture and cognition are also consistent with McNeill's (2005, p. 99) statement that gestures and speech are "not only expressions of thought, *but thought, i.e., cognitive being, itself.*"

Can gesture holds serve as automatic 'recall cues'?

Nelson and Goodmon (2003) have shown that when attention is disrupted (i.e., drawn to new material), content in working memory is frequently lost, but can be retrieved with the aid of *recall cues.* Their study was designed to investigate the effect of disrupted attention on the "impending but implicit thoughts" that people in conversation are at the point of expressing when a disruption takes place, thus their approach is highly relevant to the question of how speakers deal with interruptions during conversation. In their study, recall cues (such as words related to a target concept) were most effective when they included a reproduction of some of the context in which the target concept was first experienced. Since gestures closely contextualize speech and indeed combine with it as an integral part of utterances, gestures and other directly recruited material anchors should be particularly suited as tools for storing and representing information across periods of cognitive disruption. This is especially true of gesture *holds,* because they provide a means for a speaker's body to simply preserve some of the earlier context in which part of an utterance was being formulated or spoken: following a sudden interruption, the survival of a gesture hold means that part of that earlier context (in fact, part of that earlier utterance itself) is already in the forefront of perception and available as a recall cue. The recall cue does not need to be reconstructed; it is already there, entirely ‘free of charge’ from the point of view of a listener, and perhaps also requiring virtually no effort on the part of the speaker, if it was engaged as part of an automatic response.

The internal effects of a speaker’s own gesture holds are difficult to measure directly, but it is more straightforward to administer recall tests on a listener who has been presented with a controlled stimulus. As externalized cues, gestures are capable of serving as recall cues, in Nelson and Goodmon’s (2003) sense, for the one doing the gesturing and speaking as well as for a listener, because the gestures are part of the total “episodic experience” of both speaking and listening. Like Nelson and Goodmon, Woodall and Folger (1985) emphasized the importance of contextual concomitants for aiding recall, citing Tulving’s “encoding specificity principle” (Tulving and Thomson 1973). They found that after viewing a taped conversation, subjects who viewed silent replays of distinctive iconic gestures from the conversation were able to recall fragments of the speech the gestures occurred with, even after long delays (a few minutes, up to a whole week).[4] Thus, it appears that gestures can indeed serve as recall cues.

If gestures can serve this purpose after such long delays, it seems all the more likely that they could have an effect when still available immediately following an interruption. The notion that information can be anchored in gestural artifacts, along with the ‘recall cue’ hypothesis, allow us to combine Kendon’s (1972) suggestion about information “storage” in gesture with the idea that a later observation of such artifacts activates more than the information represented in their configurations: it also calls up other information that was being processed when the gestures were first deployed. This is exactly the sort of ability that would be advantageous for overcoming the distracting effects of interruptions during conversation. For a speaker, the embodied aspect of utterance production provides a ready-made, constantly updating material anchor to each moment of the utterance as it unfolds. Simply by freezing in place at the first sign of interference, a partial, schematic state of the utterance is preserved as if in a time capsule, with the evident ability to bring about the recall of the rest of it via a metonymic process applying to memory.

For a conversational partner engaged in utterance comprehension, meanwhile, the gestural anchors would be no less powerful as recall cues: they are closely timed with the unfolding utterance, associated with speech as part of its “episodic” context even if the speaker’s gestural meaning is not always clear. Indeed, Woodall and Folger’s (1985) results are for listeners, not speakers. This avenue of commonality between speakers and listeners, in which one person’s gestures provide stable cues lasting across intervals of time, is an additional way to unify gestures that ‘facilitate’ speaking with gestures that ‘communicate’ to addressees. There is a curiously strong tendency in the history of gesture studies to argue for one or another exclusive function of gesture, with those arguing for speech facilitation at odds with those arguing for a communicative function (De Ruiter 2003 provides a survey). But when it comes to gestures serving as material artifacts across time, both speakers and listeners experience their timing association with speech and their visual presence, even if only the speaker-gesturer retains muscular control, kinesthetic perception, and the “impending but implicit” thought processes leading to utterance production.

[4] In spite of the resemblance to Ekman and Friesen’s (1969) “emblems,” Woodall and Folger’s (1985) “emblematic cues” were actually instances of Ekman and Friesen’s “illustrator” category, not emblems.

Testing gesture holds across interruptions, experimentally

I have taken a few small steps toward experimentally investigating gesture holds as interruption-bridging recall cues (Park-Doob 2007a, 2007b), but cannot yet provide confident answers. In a pilot experiment, I recruited participants to view a series of short clips, each of which contained a brief narration of a scenario, including an iconic gesture about some object mentioned in speech. Each clip was interrupted in the middle by a visual and auditory disruption built into the video file. I tested three versions of each clip, with different participants: in the first condition, the speaker began a gesture hold shortly before the disruption, which was still visible after the disruption. In the second condition, the speaker began this same gesture hold before the disruption, but had dropped it to rest by the time the disruption was finished. In the third condition, the speaker did not gesture at all. Following each clip, participants immediately pressed a button to answer a true-false question about the clip, presented to them in audio only. Target clips, in which the question was related to content presented gesturally in two of the three conditions, were interspersed with distractors in which the question did not relate to the gestured referent.

I did not find an effect for accuracy, but one of the clips yielded significant differences in reaction time based on the three conditions: participants in the second condition (in which a gesture hold was part of the audio-video message prior to the interruption, but had been dropped to rest by the time the interruption was finished) were significantly slower to respond correctly than participants in the other two conditions. However, most of the clips produced rather large variances in response time, and failed to yield significant results. The experiment would need significant refinements before we could be sure whether the initial result was reliable, so I will not devote more space to it here.

If that result could be replicated conclusively, it would be consistent with the idea that listeners respond to having access to material anchors by partially *relying* on them: having access to seemingly stable information presented gesturally, such as a distinctively iconic gesture hold relating to one of the referents mentioned in speech, could trigger weaker internal rehearsal of that referent. Following a sudden interruption, a listener who has lost access to that gestural anchor might be negatively impacted, compared to a listener who retains access to it, or a listener who never had access to it in the first place and did not ‘offload’ any of the task of referent tracking.

Collaborative and competitive discourse with collaborative and competitive gestural artifacts

In the bulk of the previous section, I focused on ways that the attention and memory of people engaged in speaking and gesturing could be affected by their own gestures, and I did not directly illustrate with naturalistic examples how a person’s gestural artifacts could also be appropriated by conversational partners. Gestures may be deployed in conversation simultaneously for the purposes of illustration, and to help alleviate one’s own attention and memory limitations, but speakers may also find themselves holding a configuration across periods during which interlocutors attempt to assert their own vision of the affairs under discussion. This can happen when interlocutors deploy their own competing gestural representations, or it can involve an interlocutor's explicit appropriation of the original speaker's already-deployed representation. These two possibilities (which are illustrative but certainly not exhaustive) are presented below in Examples 4.4 and 4.5. Once deployed, a speaker's personal gestural anchors are no longer just personal: they become interpersonally *shared* resources that influence, and are subject to manipulation by, other people present in the interaction. This is an inescapable corollary of their artifactuality; it is one of the trade-offs of the offloading process.

These interpersonal effects were illustrated clearly by Furuyama (2000) in the context of paperless origami instruction; his research is also summarized in McNeill's (2005, p. 161) brief discussion of gesture "appropriation." My examples below are closely related, but also illustrate a tension between utilizing another's representation versus creating a competing version of one's own (for additional illustration of this tension, see Narayan 2010). In either case, configurations lasting across intervals of time are manifestations of moment-to-moment coherence, present while each speaker pauses to assess the other's speech, and to correct errors of speech or comprehension.

Example 4.4 — one person's gesture configuration as a collective resource

Many features of this example are quite evident from the images alone: Participant A deploys a gestural configuration, and her partner begins to treat the virtual locations it portrays as stable targets for her deictic gestures. The example is easily interpretable from the original video and not at all overwhelming in terms of its complexity, nor is the behavior it illustrates unusual. However, because the passage involves both overlapping speech and overlapping gestures from four different limbs, the transcript and gesture annotation become quite complex and certainly show the limitations of the scheme I have been using. Nonetheless, it is my hope that the annotations and prose description below will help the reader to glean the subtle timing details between the two speakers' speech and gestures: the claims I have been making about gesture holds in general continue to apply here.

Participant B nearly begins using a competing representation in this passage, as shown in image [b], but unlike her behavior in the next example following this one, she quickly drops it and reaches over to engage with Participant A's two stable points of reference instead. The two locations, the church and the train station (depot), are usually not 'buoyed' simultaneously by Participant A's hands, but both are implicitly present even if just one hand has an active buoy in play. This is, once again, because of metonymic, part-for-whole recognition. Her hands trade off in the 'buoy' role during the example, creating an unbroken continuity of artifactual support for the two locations, verifiably accessible by both participants throughout. For example, in images [c1] and [c2], Participant A's right hand is a buoy for the church, while she uses her left hand to attempt to refer to the details of this location. But simultaneously, Participant B's deictic hold is pinpointing the location of the depot which has just been vacated by Participant A's left hand, and which no longer has an active buoy. I discuss additional details following the example.

(4.4) **A**: So this would be the church. |**B**: Yes. ·· |**A**: Right? #

A: Then we cross the street, and this ·· would be the depot. |**B**: Uh-huh.

A: # And then [a]behind the depot, ·· there's the two trees, |**B**: 'kay behind? ·· #

⌈**A**: [b]Behi* So* [c1] # ⌉
⌊**B**: Like, ·· behind, as in does the door of the church face ···· the* ⌋

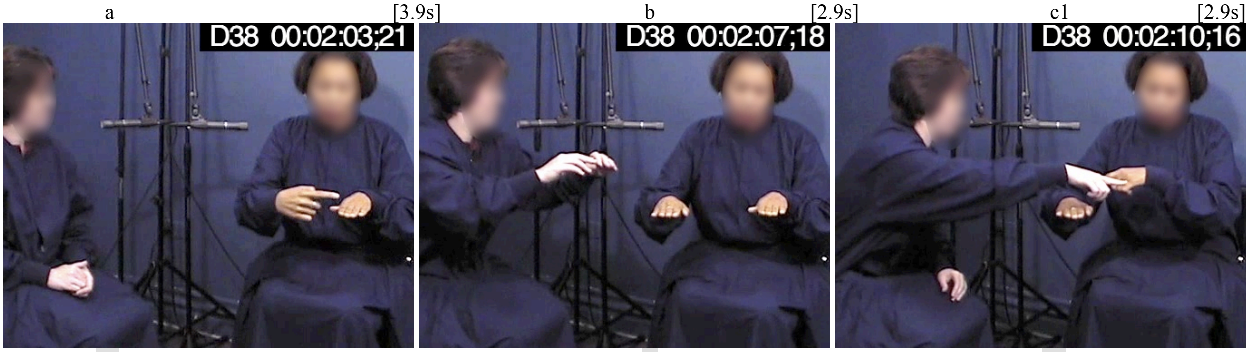
⌈**A**: So · the d* [c2] [d1] ⌉
⌊**B**: does the door of the depot face the door of the church? ⌋

A: The · door of the depot actually · faces the train tra[d2]cks.

a: *Participant A has already established the locations of the church and the depot, placing her RH as the church and her LH as the depot in a configuration identical to that shown in image* [b]. *She then leaves her LH 'depot' buoy in place while pointing to the space above it with a RH index finger deictic, shown in image* [a].

a-b: *RH index finger motions away from her with two very small beats on "behind," then holds in place at a position slightly farther away from her than in image* [a] *(not shown), after which her thumb and index finger create a momentary pinching shape with small beats on "two trees" (not shown); RH then resumes its former buoy position as the location of the church just after her partner begins interrupting (image* [b]*). As Participant B begins this interruption, she raises BH and holds them in place side by side (also image* [b]*).*

b-c1: *Participant A briefly moves her RH over to the LH 'depot' buoy again (not shown; see image* [a] *for reference), but then replaces it back into the configuration of image* [b] *just before saying "So." Immediately after her RH 'church' buoy is reestablished, she reaches over to it with a LH index finger deictic hold as shown in image* [c1]. *Meanwhile, Participant B drops LH back to rest as she reaches over with RH index finger to point and hold at the 'depot' location that was just recently occupied by Participant A's LH, as also shown in image* [c1]. *She here accidentally refers to it as the "church" but will soon correct her speech error.*



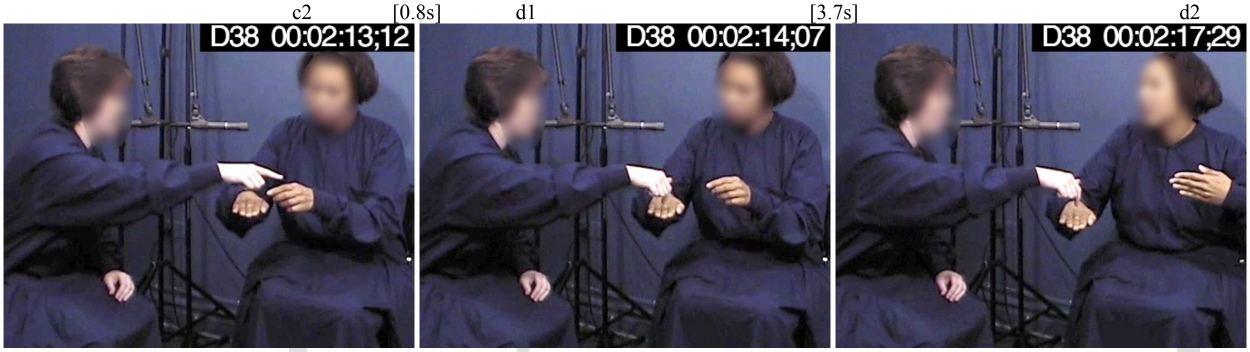
⌈**A**: then behind the depot, ·· there's the two trees, Behi* So* # ⌉
⌊**B**: 'kay behind? ·· # Like, ·· behind, as in does the door of the church face ···· the* ⌋

```
(A's RH)   ^ ^ ↗ -------- ↗        ^   ^---------- ↗ ------ ↗ ---------- ↗ ---------------------------------------
(A's LH) ------------------------------------------------------------------------ ↗      ------------------------ ↗
(B's RH) ________________________________________ ↗   -------------------------- ↗            --------^-----
(B's LH) ________________________________________ ↗   ---------------------------- ↘       ________________________
```

c1-c2: *As Participant A attempts to resume speaking, she keeps her RH 'church' buoy in place and uses her LH to attempt to pinpoint the church door, but she is interrupted and then leaves both hands held in place as her partner speaks. Meanwhile, Participant B retracts the deictic hold shown in image* [c1] *slightly, but performs it again along with her corrected speech, with a small stroke on "depot" (image* [c2]*), once again pointing to the vacant 'depot' space formerly occupied by Participant A's LH 'depot' buoy.*

c2-d1: *While Participant A continues holding both hands in place, Participant B speaks alters her index finger deictic so that it holds in place at the location of Participant A's RH 'church' buoy.*

d1-d2: *While Participant A responds, Participant B continues holding this configuration, with LH at rest and RH pinpointing Participant A's 'church' buoy. Participant A keeps her RH buoy stable while placing her LH, palm toward her body, to face the virtual location of the train tracks (image* [d2]*).*



⌈**A**: So · the d* The · door of the depot actually · faces the train tracks. ⌉

⌊**B**: does the door of the depot face the door of the church? ⌋

(A's RH) ——————————————————————————————————————

(A's LH) ∧ ———————————————— ↗ ————————————————————

(B's RH) —————∧— ↗ ——————————————————————————

(B's LH) ______________________________________

At first, Participant B attempts to refer to the location she is pointing to as the "church" (see transcript at image [c1]), but this is an error. She continues holding there during the subsequent pause in her speech, before correcting herself (see transcript at image [c2]), demonstrating the same sort of gesture hold across pauses and disfluencies that I illustrated in Chapters 2 and 3. Meanwhile, because of Participant B's erroneous use of the word "church," Participant A attempts to discuss the church's rear doors and begins pinpointing them with her left hand, as shown in image [c2]. She is quickly cut off by Participant B's corrected query regarding the depot, during which Participant A continues to hold the configuration she had assumed at the moment of interruption. As shown in image [d1], Participant B finally does refer to the church's doors that Participant A had been about to discuss, changing the target of her deictic hold to that location. Participant A then places her left hand at what would be the side of the long train station, with the flat form of her hand portraying a line or plane, as is appropriate for the shape of the train tracks (image [d2]). Participant B continues her deictic 'church door' hold during the four seconds leading up to image [d2], although she is not trying to retain the floor. Her gesture may instead be intended as a resource to be accessed by her partner as she responds, a phenomenon I discuss in the next chapter.

Example 4.5 — maintaining gestural configurations against competing alternatives

In this example, taking place less than a minute after the exchange illustrated in Example 4.4 above, Participant B makes no attempt to utilize her partner's gestural placements of landmarks, instead choosing to set up her own. But Participant A does not cede her own placements; instead she retains them in competition with Participant B's gestural map, to be ready for the inevitable clarifications she must make in response to the latter's repeated requests for additional spatial detail. As a result, three simultaneous gestural 'buoys' (two in competition for the depot, and one for the church) are kept in play continuously across speech by both participants (see images [d1] through [h]).

(4.5) **A**: We're supposed to get ····· from Houses number 33 · which is like the far house.

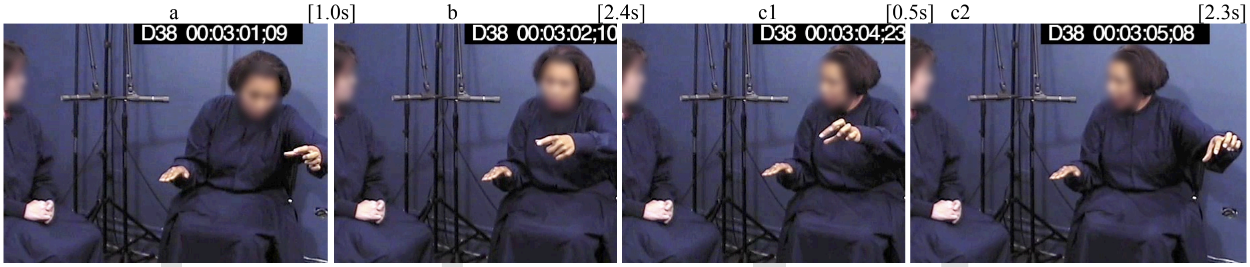
```
      a          b                         c1       c2
[A:  33,   ·   and 35. ····· Right? There's · two little houses.                 ]
[B:  Uh-huh.              Now,              how are they far from the h* ·· from the depot?]

                d1               d2                d3                    d4
[A:   They're not ·· really that far.  ···                               Uh-huh?  · ]
[B:  Are     they*     #              N'okay but · here's the depot,  ·             ]

          e         f        g          h
 B:  The trees, the rocks, · house and house?
```

a: *Leading up to image* [a], *Participant A's RH 'church' buoy has survived the intervening 43 seconds since the last image of Example 4.4 above (albeit with a few brief retractions after which it was quickly put back in place). She then uses her LH in a thumb and forefinger 'pinch' shape to place House 33, shown in image* [a].

a-b: *While her RH buoy is maintained, she uses the same LH handshape to place House 35 about one foot to the side of House 33. Her L arm is almost fully extended during these placements, while her R arm is fully bent at the elbow, creating a large distance between the 'church' and the houses. This reflects the real configuration of the absent model village.*

b-c1: *Her LH twists at the wrist; she then performs a downward stroke at the 'House 35' location.*

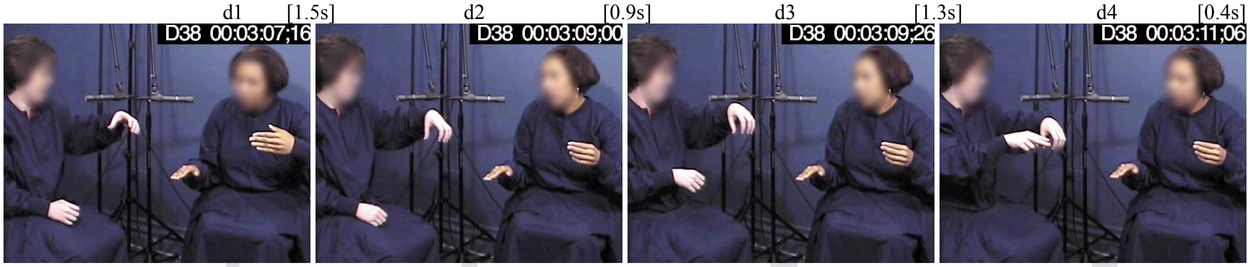
c1-c2: *Her LH sweeps back to her far left and performs another downward placement stroke, this time at the 'House 33' location.*



```
[A:        Thirty-three,    ·    and    Thirty-five. ····· Right? There's · two      little     houses.                  ]
[B:              Uh-huh.                                  Now,              how      are      they far from the h* ·· from the ]
(A's RH) ——————————————————————————————————————————————————————————————————————————————————
(A's LH) ——  ^ ——————————— ↗     ^ ——————————— ↗        ——— ^            ^       ———————————————————
```

c2-d1: *As she answers her partner's interrupting question, Participant A keeps her RH 'church' buoy constantly in place and moves her LH palm toward her body, indicating the position of the depot relatively close to the houses. Meanwhile, at the same time that Participant A moves her LH to show the depot's location, Participant B raises her own LH to independently maintain her own placement of the depot. In the final eight images of the example (images* [d1] *through* [h]*), two separate 'depot' buoys are maintained, both by a participant's LH.*

d1-d4: *Both participants maintain these placements (the 'church' by Participant A's RH, the 'depot' by her LH, and the 'depot' independently by Participant B's LH), then Participant B partially raises her RH but holds it in place near her lap as she executes a downward beat with her 'depot' LH at the moment shown in image* [d3]. *She then brings her RH, two fingers extended, to unify it with her LH's location and pause for a moment (image* [d4]*).*



⌈**A**: They're not ·· really that far. ··· Uh-huh? · ⌉
⌊**B**: depot? Are they* # N'okay but · here's the depot, · ⌋

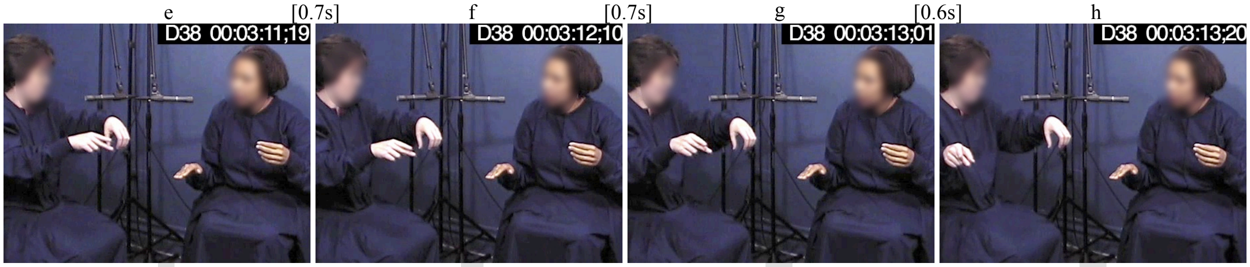
(A's RH) ——————————————————————————————

(A's LH) ——↗ ——————————————————————————

(B's RH) ——————————— ↗ ——————— ↗ ——— ↗

(B's LH) _ ↗ ——————————— ^ ———————————

d4-h: *While BH of Participant A and the LH of Participant B maintain their configuration, Participant B uses her RH to pinpoint a series of locations in relation to her LH 'depot' buoy. She lifts her RH slightly between each while moving it outward in a straight line, performing a downward stroke at each location shown in images* [e] *through* [h].



B: The trees, the rocks, · house and house?

(A's RH) ——————————————————————————————

(A's LH) ——————————————————————————————

(B's RH) ^ ↗ ^ ———— ↗ ^ — ↗ ^ ———

(B's LH) ——————————————————————————————

At the start of the example, Participant A's right hand 'church' buoy is already in place, as it has been nearly continuously since the previous example (later in their conversation it exhibits even greater longevity, as was illustrated above as Example 4.3). She uses her left hand to place the two houses, first No. 33 and then No. 35 as shown in images [a] and [b], and then reiterates their locations in the reverse order with a sideways sweep of her hand as shown in images [c1] and [c2]. During this reiteration she resists Participant B's interruption and completes the utterance, after which image [d1] shows clearly the moment at which their competing representations begin: as shown in the gesture annotation below the image, both participants simultaneously move their left hands to begin representing the depot's location. From image [d1] to [d2], Participant A has interrupted Participant B's query, and the latter keeps her left hand (thus far unidentified) in place as she waits to repeat her query in earnest. At image [d3], she has begun raising her right hand but pauses after a partial preparation in order to perform a downward movement with her left hand to explicitly identify it with speech as the depot. The final four images then show Participant B's right-handed placement of the trees, road construction, and houses in relation to her left hand 'depot' buoy.

Participant A, meanwhile, retains her own configuration in order to be ready to address her partner's queries without having to utilize the latter's frequently erroneous representations. Her own locations are stored as gestural artifacts, so she can afford to direct her attention to tasks such as the assessment of her partner's version, without risk of losing track of her own. Any degradation of her own configuration in working memory would be quickly restored upon experiencing her own gesture holds as recall cues (see the discussion in the previous section of this chapter).

Example 4.6 — coordinating simultaneous timelines of collaboration and competition

This chapter's final example illustrates a rather different angle on the issue of cooperative and competitive speech and gesture in conversation. This time, although the participants' speech overlaps significantly, it is also explicitly collaborative, with each assessing and responding to the other's ideas. At the same time, however, Participant B has additional details in mind, which his partner does not immediately take up and which he therefore wishes to reiterate each time he regains temporary control of the floor. He carries through evidence of these impending reiterations in the form of his gestures, lasting across his partner's speech turns, and even lasting across his own separately collaborative speech dealing with different content from what the form of his gesture is maintaining. In so doing, his gestures in effect maintain his competitive ambitions on one tier, while he remains explicitly collaborative in speech. The effect is illustrated with his circular 'net' outline in images [a] through [c4], and even more strikingly with his 'cage' handshape in images [d] through [e4]. Across his partner's contributions and his own collaborative remarks, his gestures maintain a link to previously mentioned ideas that he is about to reiterate, and their survival through these insertions provides a ready-made defense against any loss in memory of those speaking plans. Intermodally, these separate goals play out on differing timelines, yet speech-gesture synchrony is actually maintained due to a complex layering phenomenon, which I discuss following the example.

(4.6) **B**: I say we have all these locals out here with all the nets and the tasers? #

a b1 b2 b3
[**A**: Yeah. Oh yeah.]
[**B**: We should like* · so have a like · a circle, surrounding this area.]

b4 c1
[**A**: We need to. · Definitely. We have a defense perimeter right there.]
[**B**: Like · a net.]

c2 c3
[**A**: A net oh that's a good idea.]
[**B**: But we should have like · a net · right here.]

c4 d
B: And then we have all the people inside with like,

e1
[**A**: Alright, so we're gonna need to bring a big ass net.]
[**B**: like cages.]

e2 e3
B: A big ass net.

A: We gotta remember that.

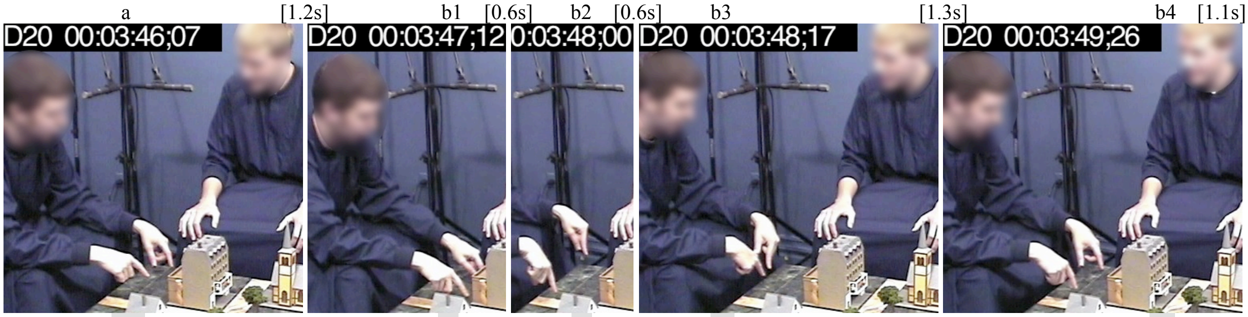
e4 f g h
B: With cages, ·· and then they're gonna have tasers just in case things get outta control.

a: *In the lead-up before image* [a], *BH indicated the houses (timed with "all these locals"), and then moved to his left to form a symmetrical spread-fingered, loose dome shape over the rear of the theater (timed with "here," and then held there until through to image* [a]*). He then leaves index fingers extended but curls his other fingers in, as shown in image* [a].

a-b1: *BH tips of index fingers thrust forward to almost touch the rear of the plastic theater model.*

b1-b3: *BH are then drawn back toward him in a slow arc, with the fingers coming together to close the circle at image* [b3] *just after Participant A has begun overlapping his speech.*

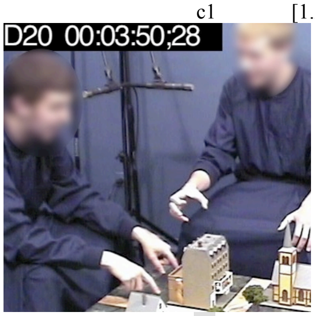
b3-b4: *As Participant A speaks, Participant B keeps BH index fingers pointing down and, at an extremely slow rate, begins retracing them forwards along the circular path they have just completed.*

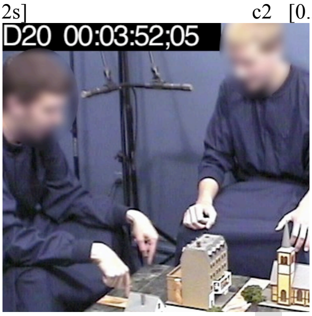


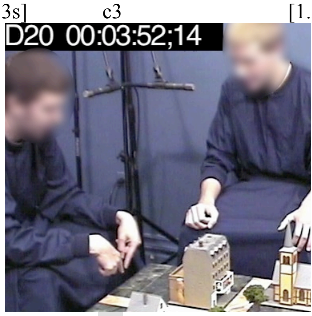
[**A**: Yeah. Oh yeah. We need to. · Definitely. We have a defense]
[**B**: We should like* · so have a like · a circle, surrounding this area.]
—— ↗ ——————————^——————^—— ———— ——————————————————————

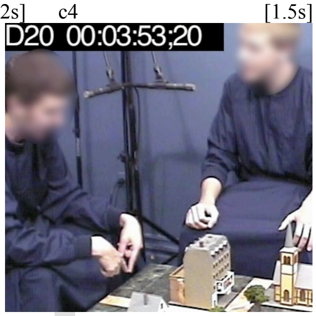
b4-c1: *BH fingertips reach back of theater, then each spreads to sides slightly, timed with "net."*
c1-c2: *BH hold there, and RH finger taps with two subtle beats, then BH sweep the circle again on "net."*
c2-c4: *BH fingertips come together again, on "here," and hold in place through Participant A's speech.*

c1 [1.2s] c2 [0.3s] c3 [1.2s] c4 [1.5s]





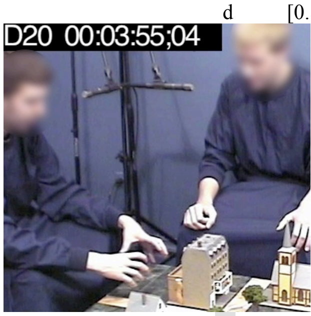


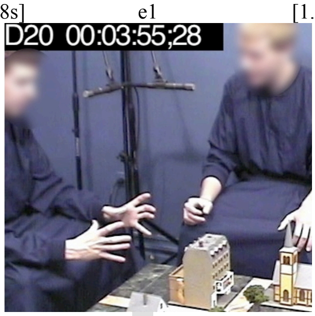


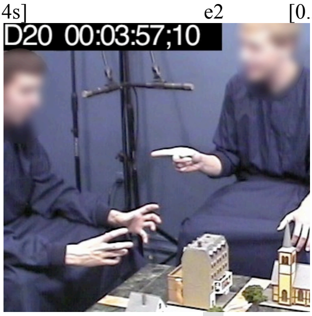
⎡**A**: perimeter right there. A net oh that's a good idea. ⎤
⎣**B**: Like · a net. But we should have like · a net · right here. And then we have all the ⎦
––––––––––––––^––––––––––––^–––––––––^––– ––––––––––––––––––––––––↗ ––––––––––––

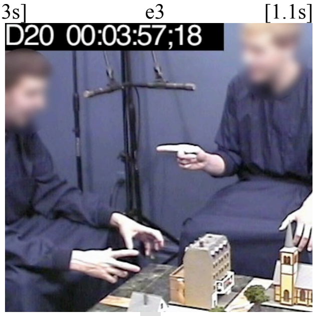
c4-d: *Exactly as Participant B resumes speaking (after image* [c4]*), he sends fingers forward to rear of theater again (not shown), after which BH immediately spread into claw shapes (image* [d]*).*
d-e2: *BH raise and hold in more pronounced claw shapes as partner speaks, with RH beats on "cages."*
e2-e3: *As he repeats "big ass net," BH thrust in a short downstroke, but continue holding claw shapes.*
(e1-e4): *(Partner's deictic starts on "gonna"; stroke with his "big ass"; waggles + retracts on "We gotta").*

d [0.8s] e1 [1.4s] e2 [0.3s] e3 [1.1s]





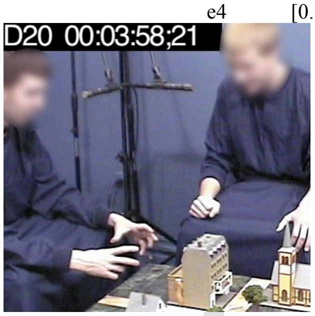


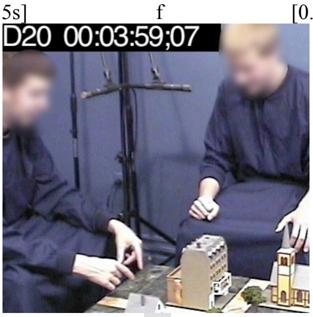


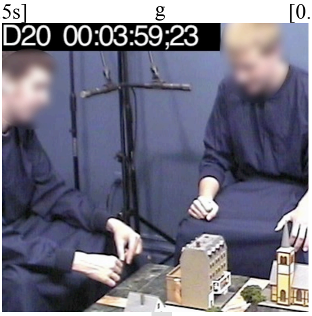
⎡**A**: Alright, so we're gonna need to bring a big ass net. We gotta ⎤
⎣**B**: people inside with like, like cages. A big ass net. ⎦
–––––––^–↗ –––––––↗ ––––––––––––––––^–^––––––––––––––––^ –––––––––––––––––––––––––––––––

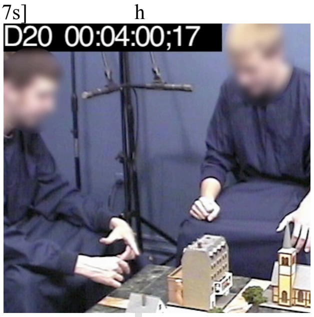
e4-h: *BH hold until "with cages," then become points (*[f]*), then retract (*[g]*) prior to new gesturing (*[h]*).*

e4 [0.5s] f [0.5s] g [0.7s] h









⎡**A**: remember that. ⎤
⎣**B**: With cages, ·· and then they're gonna have tasers just in case ⎦
–––––––––––––↗ –––––––––––––––↘ ______________________ ^ ––––––––––––

After Participant B performs the ‘circle’ gesture at the beginning of the example (images [a] through [b3]), his partner immediately responds in speech, overlapping the end of his utterance. As Participant A speaks, Participant B’s intention to elaborate on his earlier statement is evident (to us, the analysts, but not necessarily to his partner) from the fact that he very slowly retraces his index fingers along the circular path while his partner speaks. Just before the moment shown in image [c1], he has assessed his partner’s “defense perimeter” suggestion and overlaps the latter’s speech to reassert what he really intends the “circle” to be: the outline of a net surrounding the area. He emphasizes this by performing the gesture again with his revised speech (see images [c1] through [c3]), then holds its ending configuration through his partner’s speech turn (images [c3] through [c4]), because he wishes to continue speaking.

He then switches to a new idea, which is that there should be people manning cages inside the netted perimeter. His ‘cage’ handshape can be seen taking form in image [d], reaching its full extent in images [e1] and [e2], with small emphasizing beats on “cages.” However, during the time leading up to his mention of “cages,” his partner has already jumped in to emphasize (along with an index finger deictic directed at Participant B) that they must remember to bring “a big ass net.” In both speech and gesture, Participant B cooperates, repeating his partner’s words along with an emphasizing downward stroke of his hands. Crucially, however, he retains the ‘cage’ *handshape,* even during his gesture stroke emphasizing “a big ass net,” and after another turn by his partner he is able to mention the cages again. He retains the ‘cage’ handshape until just before mentioning the referent again in speech, at which point he uses his index fingers to point to the interior area where they would be placed (image [f]).

During this episode, Participant B retains speech-gesture synchrony, but the intermodal timing is complex: his *stroke timing* lines up with prosodic emphasis of the speech regarding the “big ass net,” allowing him to superimpose collaborative speech and gestural emphasis onto a pre-existing and unrelated *handshape.* The handshape, meanwhile, maintains reference to a previously mentioned idea that he is about to reiterate, and its survival through his collaborative digression provides a ready-made defense against any loss in memory of those speaking plans. In the process of emphasizing his agreement with his partner’s “net” idea, he may essentially cancel out the ability of his ‘cage’ gesture to communicate its originally intended content. Yet the gesture can live on as a manifestation of his unrealized speaking plans, carried through as an embodied recall cue even if completely ignored by his partner. In terms of holds and strokes, his speech-gesture timing is both collaborative and competitive—collaborative in terms of the *instantaneous* timing of the stroke with overt speech, yet implicitly competitive in the shape of the hold, a gestural artifact which is *retrospective* of speaking plans that were not accomplished to his satisfaction and *prospective* of his intended reiteration.[5]

[5] This example describes a gesture accomplishing two things at once because of a stroke embedded within a held handshape—that is, it relies on the differing *timing* characteristics of strokes versus holds, which allow a *transient* movement to be superimposed on a *lasting* shape. This is akin to the phenomenon of beats superimposed on holds, as discussed by McCullough (2005), except that those beats were characterized as supporting the same reference as the hold rather than accomplishing a separate goal. But for additional discussion of single gestures accomplishing two things at once, in this case by way of a shape representing one physical *viewpoint* in a description while a movement trajectory represents another, see Parrill 2009.

Conclusion

In this chapter I first emphasized that many of the same features of gesture holds coinciding with speech disfluencies also occur with gestures spanning interruptions. While previous chapters focused on the utility of such gestures for allowing a listener to reconstruct complete utterances from fragments, the current chapter focused more on the capacity for such gestures to store, 'retrospectively' but for future use, a speaker's plans when faced with disruptive interference from interlocutors.

Because of constraints on the functioning of attention and memory, there may be an inescapable drive to rely on physical anchors as 'recall cues', especially when their form maintains part of the state of an utterance or utterance context. In this way, a gesture hold can form a 'bridge' across moments during which attention may be forcibly drawn elsewhere. Of contextual features which last across intervals of time, a gestural artifact is among the most closely integrated with the process of utterance, and can also remain stable across moments of distraction, creating a recall cue that is automatically present and salient in perception following disruptive events.

When long-lasting gesture holds are deployed, they become susceptible to appropriation, and can also end up in competition with another's gestures vying to anchor the same referents in different arrangements. These are unavoidable consequences of using artifactual gestures in the service of cognitive and communicative tasks. But gesture holds can also embody and retain a speaker's competing, partly unrealized utterance plans on one timeline, even as speech and superimposed gestural movements reflect a simultaneous timeline of explicit collaboration. In the longevity of the gesture's form, a planned utterance move can be kept on hold while a speaker responds instantly to interpersonal cooperative expectations.

5 Gesture holds maintained across turn transitions

In nearly all of the examples presented so far in these chapters, gesture holds have coincided with an intention to continue speaking. Indeed, Gullberg and Kita (2009) have noted (citing earlier research going back to Duncan 1973) that turn transitions are more likely upon termination of a gesture (i.e., retraction) or relaxation of a tensed hand position than they are during gesture holds, meaning that holds may be able to help speakers maintain control of the floor by inhibiting speech from interlocutors. Consistent with that account, I argued for the utility of an 'automatic hold response' during speaker disfluency, as well as for the utility of holds lasting across pauses in speech during preliminary, incomplete commitments to utterance (see the discussion in Chapter 3 of the 'commit-and restore' model of Clark and Wasow 1998). In addition to the capacity to bear significant utterance content and make it available even during disfluent speech, such gesture holds can perform the independent function of discouraging interruptive contributions by interlocutors. The retraction of a hold then marks a relinquishing of control and, in some cases, an invitation for response, which I argued was grounds for including gestural considerations in any account of how responses are 'mobilized' (see Stivers and Rossano 2010).

In light of the above, it is therefore worth exploring why some gesture holds are maintained when speakers are in fact seeking to yield the floor and obtain a response, a behavior seemingly at odds with the functions just outlined. This is a separate role for holds than what I described in Chapter 4, which illustrated the phenomenon of speakers maintaining gestures across interruptions by interlocutors. In that set of examples, the gestures are more akin to holds coinciding with disfluencies, because even though they span turn transitions, the 'interruption-bridging' holds are still consistent with *attempting* to maintain the floor and inhibit interlocutor speech. This is in addition to the ability of these holds to act as temporary gestural artifacts and preserve aspects of suspended utterances across periods of redirected attention and memory decay.

In contrast, the examples in the current chapter illustrate holds that are maintained beyond the point of a speaker's successful utterance completion, and into the start of an interlocutor's response. These holds, which I dub 'handoffs', are then usually released soon after the response begins. They are distinct from extended gesture 'buoys'—gestures which are similarly maintained beyond the successful completion of the utterances in which they emerge—because the latter are, once again, consistent with a goal of continued speech by the gesturer rather than the release of the floor while seeking a response. The ‘handoff’ holds I will discuss are also distinct from a type of response-seeking gesture described by Kendon (2004, pp. 130-132), in which a change to a new gesture phrase (that is, a major change in gesture form) coincides with the seeking of response. His example is of a separable gesture form embodying a “kinesically held question” (Bavelas 1994, p. 203, attributed to Kendon) but not necessarily embodying a referential function, whereas the examples I will discuss are gestures used for making reference during an utterance but which are then held in place while awaiting a response. The ‘handoff’ gestures I illustrate are therefore akin to ‘buoys’ in one respect, which is that they are kept in play as a relevant cue for elaboration by additional speech; the difference is that they are maintained *for use by the listener,* and released soon after the listener becomes the new speaker.

‘Handoff’ holds: Enforcing context while transferring control

In terms of the gesture’s form and timing leading up to the turn transition, there is often nothing to distinguish a ‘handoff’ hold from one coinciding with a pause during which a speaker intends to continue without seeking a response: both involve a gesture that was synchronized with speech and whose continued presence is therefore ‘retrospective’, and both last beyond the end of a *completed* (rather than interrupted or abandoned) ‘tone unit’ of speech (see Kendon 1980). In either case, therefore, the lexicosyntactic cues are often consistent with a complete turn at talk (see De Ruiter et al. 2006), and indeed, I have argued that one of the functions of holds can be to maintain control of subsequent speech even when lexicosyntactic cues indicate an opportunity for turn transition. It is therefore curious that these ‘handoff’ holds seem to coincide with an intention to relinquish control: the gesture form itself is not signaling this, so we must look for it elsewhere.

One obvious way in which a response can be sought, in spite of a persistent gesture hold, is when the speech coinciding with the gesture is interrogative. However, as I illustrate below, some holds intended as ‘handoffs’ occur with non-interrogative speech, so there must be another way to initiate the same transition, independent of both speech and manual gesture. This key behavior is listener-directed gaze. As discussed in Chapter 2 (see especially p. 42), Stivers and Rossano (2010) have found that gaze directed at interlocutors is one of the most important behaviors capable of ‘mobilizing response’ (interrogative speech is another). Their work builds on Kendon’s early (1967) account, which is also quite consistent with the examples I illustrated in Chapters 2 and 3: in short, speakers often glance at their interlocutors when ready to yield the floor, and tend to look away when they begin a turn at speaking as well as when they hesitate or become disfluent.

When looking at a listener while finishing an utterance (or after finishing it), and with continued maintenance of a hold, the gaze cue evidently overrides any pressure against speaking that the listener may feel due to the speaker’s continued gesture. The hold is an artifact of the speaker’s prior utterance, normally providing evidence of the speaker’s intent to elaborate, but it instead becomes the basis for elaboration by a new speaker. The new intended speaker is thrust into the role because of the previous speaker's directed gaze, which indicates an assumption of the role of listener as well as the target of listening. Nothing in this mechanism requires the initiation of a new gesture phrase: the hold remains a 'retrospective' extension of an earlier utterance, and it is only the expected next speaker that has changed. Yet the earlier utterance's extension also ensures that the previous speaker enjoys some measure of control over the new speaker's utterance, because the eyes demand a response while the gesture enforces a context.

The examples below illustrate variations on this theme. In the first, the transitions between the participants' turns occur by way of a chain of alternating handoffs, with each gesture held just until the next speaker begins. The second example demonstrates a partial gesture retraction at the moment of handoff, which seems to reflect partial, but not total, relinquishing of control. The third and final example demonstrates additional handoffs, as well as cases of listener-directed gaze combined with gesture holds that *persist* even when the listener becomes the new speaker. This last combination of behaviors allows the original speaker to gain a response while also maintaining control and broadcasting an intention to resume speaking.

Example 5.1 — a 'handoff' chain between alternating speakers

Although the faces of the participants are blurred, in most cases it is possible to distinguish the angle of their heads relatively well, which closely matches the direction of gaze in the images I have selected. For example, in the first row of images below, Participant A, seated on the right, is gazing at her partner in three out of four images. Image [b1] is distinct from the others in that she has tilted her head—and her gaze—down to the rear of the model movie theater.

The first of three handoffs in this example coincides with listener-directed gaze, the second coincides with explicitly interrogative syntax, and the third coincides with interrogative intonation. According to the account by Stivers and Rossano (2010), both gaze and interrogative speech are, along with other factors, independently capable of indicating to an interlocutor that a response is required or appropriate. They frequently co-occur, but an utterance containing just one or the other can still 'mobilize response', and a gesture hold being maintained past the end of the speaker's turn will in either case be understood as relevant for the response.[1]

The first handoff begins around the moment of image [b3]. Participant A has finished the spoken part of her utterance but keeps her finger pointing at the back of the model theater, while gazing at her partner. She maintains this configuration through one second of silence, after which her partner realizes she must respond. Just before the

[1] The emphasis on the gesture can vary, of course. Given that listeners appear to attend most conscientiously to gestures that are being gazed at (Gullberg and Kita 2009), the use of explicitly interrogative forms in speech can allow speakers to seek a response while gazing at their own gestures (or their deictic targets), which should emphasize their relevance—this gaze-marking of communicative relevance was a major claim of Streeck's (1993). Alternatively, a speaker using interrogative forms in speech could elect to emphasize the gesture somewhat less, while more vigorously seeking a response, by gazing at the listener instead.

moment of image [b4], Participant B offers a preliminary "okay," then formulates a further response while Participant A continues to hold the gesture and gaze at her. Between images [b4] and [c1], the participants switch roles: Participant A retracts her gesture just as Participant B is initiating a full turn at talk, which turns out to be a new query coupled with its own sustained gesture (note that during this transition, Participant B's right hand has become independently occupied with the task of adjusting her glasses—it is only her left hand's deictic gesture that is a part of her utterance). Participant B maintains her deictic during the pause following her spoken query, allowing it to continue targeting the referent in question as Participant A looks at it and forms a response. Finally, just after the moment shown in image [c2], the participants switch roles again: all at once, Participant B retracts her gesture, and Participant A begins speaking while initiating a her own new gesture. She maintains it as a hold during Participant B's brief response (image [d]), then retracts it.

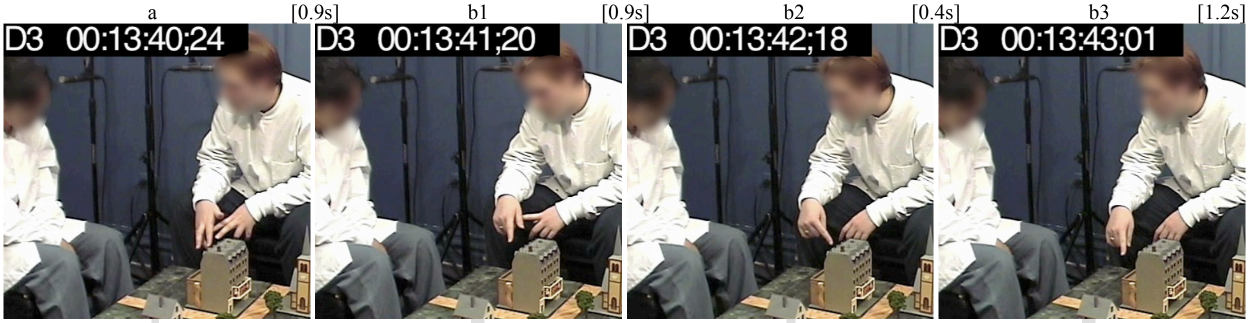
(5.1) **B**: I'll climb up the ladder and · see where they are and then climb back down.

A: Okay. Do you think we'd be better to just do·· * open one door?
# Maybe if you can find out where they are ···
we won't · need to do ··· both doors, ·
a b1 b2 b3
if they are closer to one side or the other, ·····

b4 c1 c2 d
⎡**A**: Okay. Yeah on the front? · ⎤
⎣**B**: Okay. ·· Is it* · Are there doors on the other side? ·· M-hm. · ⎦

A: Yeah. ··

a: *Participant A has fingers of BH extended in a close deictic hold over the rear area of the theater, and is gazing at Participant B.*
a-b1: *Participant A directs her gaze at the rear of the theater, while relaxing LH and extending RH index finger to point at the doors at the rear of the theater.*
b1-b3: *She returns her gaze to Participant B's face, while her RH index finger points from side to side in a repeating, alternating pattern (see the stroke marks in the gesture annotation below).*



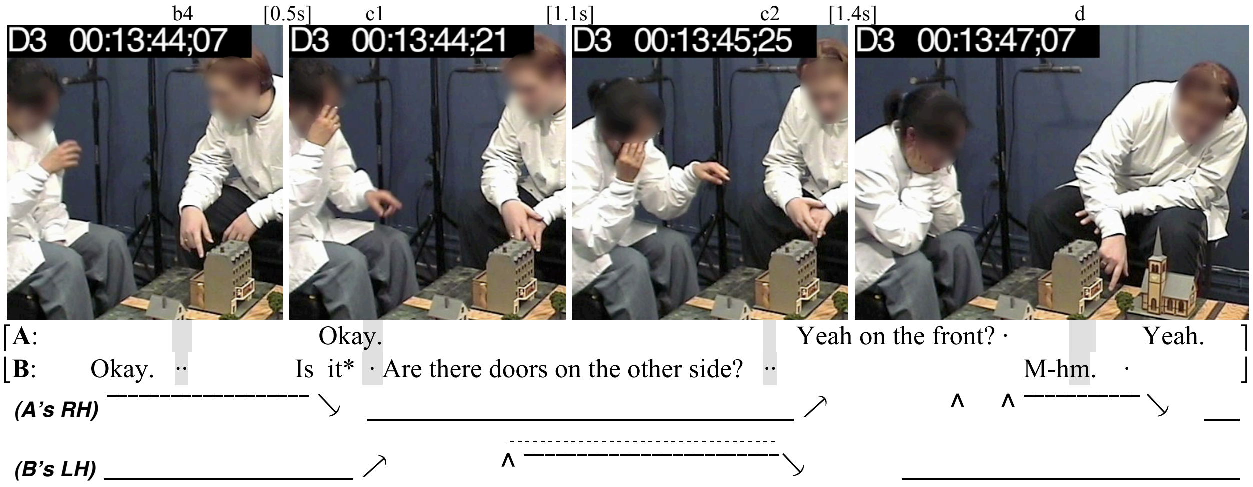
A: if they are closer to one side or the other, ·····
(A's RH) ———————— ↗ ^ ^ ^ ^ ^ ^————————

b3-b4: *Participant A continues to gaze at her partner during one second of silence, while also holding her RH deictic in place (it relaxes slightly). Participant B begins to raise RH to adjust glasses (this is not coded in the gesture annotation – it is Participant B's LH that will be relevant).*

b4-c1: *Just after the moment of image* [b4], *at* [44;09], *which is 0.6 sec. after Participant B says "Okay," Participant A averts her gaze and retracts her hold exactly at the moment she also says "okay." About 0.2 sec. later, at* [44;15], *Participant B begins raising her LH from between her legs.*

c1-c2: *Participant B's LH becomes an index finger deictic pointing over the top of the theater, which she waggles in a jabbing manner while continuing to gaze at the model until just after the moment shown in image* [c2], *and Participant A turns her head to look there.*

c2-d: *At* [45;29] *several things happen almost simultaneously in the span of three frames (0.1 sec.): Participant B begins retracting LH to chin, and Participant A begins a new verbal response-initiative and extends RH for a new index finger deictic toward the front of the theater, holding there after a double stroke on the word "front." She maintains the hold while her partner responds, then retracts it to rest as she says "Yeah."*



⎡**A**: Okay. Yeah on the front? · Yeah. ⎤
⎣**B**: Okay. ·· Is it* · Are there doors on the other side? ·· M-hm. · ⎦

The collaborative timing of speech and gesture in this example results in a closely coordinated process that plays out almost as if it were choreographed: each speaker's gesture is maintained just long enough to last into the beginning of the next speaker's turn, then quickly retracted. As shown above in the gesture annotation below images [b4] through [d] above, the flow of the participants' 'hold' versus 'rest' positions proceeds in striking mirror-image fashion, with each participant fully relinquishing the speaker role as the other responds.

Example 5.2 — a 'handoff' with graded retraction

I first discussed the beginning of this example in Chapter 3 (Example 3.3, pp. 80-82), with regard to the gesture hold spanning Participant B's speech repair. This time, my focus is the longevity of the hold, which she maintains beyond the end of her query and into the start of her partner's turn. The target of the deictic is the train station (which Participant B is interpreting as some kind of park building) and the hold keeps this referent relevant as the desired focus of her partner's response. When this response begins, Participant B relinquishes control with a retraction, but it turns out to be a *partial* retraction: she lowers her arm to her lap, but retains the pointing handshape and the same deictic target. This occurs just after the moment shown in image [b3], at which point she appears to simultaneously yield the floor while also retaining some outward evidence that she still wishes to speak about the referent. Indeed, following her partner's identification of the landmark, she speaks again while still maintaining the hold (see images [b4] and [b5]), then fully retracts her gesture as she finishes the utterance (see image [c]), satisfied that she has finished considering the referent and will not continue speaking about it.

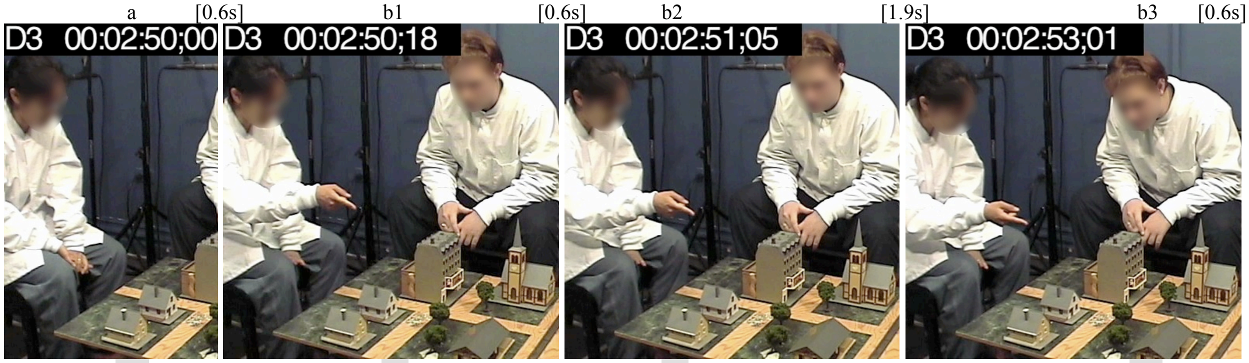
(5.2) **B**: So in the directions did you receive · information about any other people in this town?

a b1 b2
[**A**: # · No.]
[**B**: Like · does nobody work · in · the ·· what is that the park? ···]

b3 b4 b5 c
[**A**: That's · that's a train station? # Uhm ··· but]
[**B**: Oh the train station. · so nobody'll work there.]

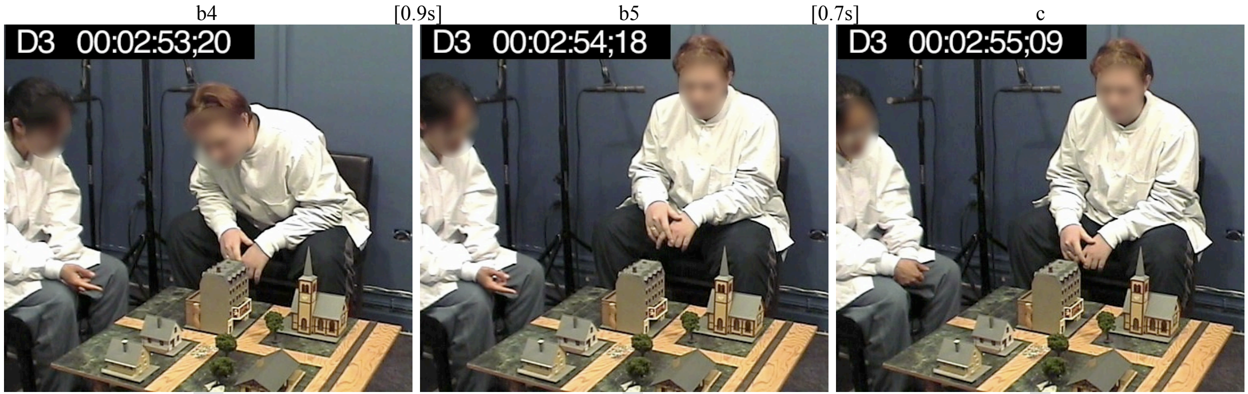
A: they're ··· busy with · train station type things.

a-b1: *While Participant B gazes at the model, her RH rises from her lap to form an index finger deictic hold that targets the trees and building visible at the lower right of the images.*
b1-b2: *Less than 400 milliseconds after her partner's deictic takes form, Participant A turns her head almost imperceptibly to her left to attend to the target of her partner's gesture.*
b2-b3: *As her partner maintains the hold, Participant A leans forward toward the mutually attended referent and responds while raising her eyebrows.*



[**A**: # · No. That's · that's a]
[**B**: Like · does nobody work · in · the · what is that the park? ···]
(B's RH) ___________ ↗ ^ ————————————————————————————————

b3-b4: *Just after Participant A completes the word "that's" in speech, Participant B's arm begins dropping to her thigh, but although her limb returns to rest she retains the deictic handshape.*
b4-b5: *Participant B continues to maintain the deictic during the first part of her response.*
b5-c: *As she completes her response she fully retracts her index finger deictic handshape.*



A: train station? # Uhm ··· but they're ···
B: Oh the train station. · so nobody'll work there.

It should be noted that Participant B is not necessarily making any conscious separation of the task of orienting herself to the referent, versus getting her partner to respond to her deictic target. The referent is an object of *mutual orientation,* intended for collaborative consideration, as achieved partly through gesture (see Goodwin 1986). While she is still puzzling about the referent, she maintains some of the gesture, and only fully retracts it when she is satisfied that she understands the landmark's identity. The gesture's longevity is evidence of her continued thought, and the gesture's outward presence makes it appropriate as a cue for orienting Participant A's response (a response which is itself a tool recruited in the service of Participant B's goals). Just as the long-lasting gestures in Example 4.6 (pp. 119-122) coincided with the speaker's continued pursuit of certain speaking goals across his partner's competing contributions, the partially retracted hold in the example above is an outward manifestation of Participant B's continuing engagement with the referent.

Example 5.3 — holds released via 'handoff' versus holds maintaining control

This final example contains multiple instances of holds lasting across turn transitions, demonstrating a useful variety of behavior that will help to link the discussion back to previous chapters. In every instance below, the speaker is directing gaze at his listener and his speech is usually *not* interrogative. The timing of his gesture holds places them into two major categories: those which he retracts once his partner begins responding, and those which he keeps held in place. In the latter combination, his gaze allows him to secure a contribution from his partner, which he intends to be contingent on the referents indicated by his continued gesture, but his lack of retraction also indicates a verifiable intention *not* to yield the floor.

The first row of images below illustrates a relatively typical lifecycle of a 'handoff' hold. The speaker points at a target (in this case, the model church) and begins looking at his listener. In the middle of the utterance, he looks back at the church while adjusting his gesture to a closer, more specific deictic, and immediately resumes gazing at his partner. By the time the participants reach the moment shown in image [b2], the speaker has paused and then added an interrogative in speech, while continuing to maintain the hold as well as his listener-directed gaze. Crucially, although his speech contains an interrogative and this indicates convincingly that he is seeking a response, his partner has already begun to respond by the time the interrogative is spoken, so it cannot be claimed that the interrogative had a role in mobilizing the partner's response in this instance. Approximately one half-second after Participant B begins responding, Participant A retracts the gesture back to rest (see image [c]).

The rest of the images of the example are a continuous set, but note that they begin ten seconds after the moment shown in image [c], picking up while Participant A is already engaged in a hold directed at the small trees on the 'grass' behind the train station. Images [e1] and [e2] show another response-seeking process, in which he turns his head toward his partner just after the moment shown in image [e1], then reiterates the stroke and holds the gesture in place, waiting for some kind of response. This time, his partner offers only a backchannel acknowledgment, and Participant A does not relinquish control. Instead, he begins a new utterance with a new gesture preparation. Although it is impossible to determine with certainty, it is plausible that Participant A would have retracted his gesture fully if his partner had launched into a full turn. Instead, with his partner's very brief and slightly delayed response he evidently feels compelled to seek further verification, with a closer, more carefully directed gesture. The contrast between images [e2] and [f1] may be revealing of different goals: the first could be merely a step in the narration, such that Participant B's backchannel response is all that is expected. Participant A then elects to 'drive home the point', as it were, and require his partner to take careful note and demonstrate his uptake explicitly. This demand may be a result of the fact that Participant A had waited in silence for nearly a half-second before his partner finally said "Uh huh," which likely cast doubt on the latter's uptake.

From [f2] to [g], Participant A's deictic hold again behaves as a 'handoff': his partner points at the house and repeats its number in speech, during which Participant A immediately retracts his gesture while repeating the number once again (in a softer voice). However, Participant B simultaneously attempts to display further understanding by pointing at the next house and stating its number as well, incorrectly, which causes A to spring back into action before he has fully reached rest position. His final hold (images [h1] and [h2]) is directed at the previously unnamed House 35, adjacent to the theater, and it is *not* a handoff in the sense of seeking a response during which the floor will be relinquished, as I will discuss following the example.

(5.3) **A**: So ····· We gotta go, ·· take the train in, ·· go to the train station,

and we'll kn[a]ow it because · # [b1]the church is right before it. ··

[**A**: Right? [b2] [c]]
[**B**: Okay well I mean yeah I assume the train will stop at the train station right? ······]

[**A**: I hope so. Anyway@ # ··]
[**B**: @ @]

A: Umm, · so then we get off, ·· and we go through this little park right here. ··

Cut between these t* · these [d]trees and go to [e1]House Thirty-three. ···

[**A**: [e2] Thirty-three. [f1] [f2] Thirty - three.[g]]
[**B**: Uh-huh. · Thirty-three, Thirty-two. ··]

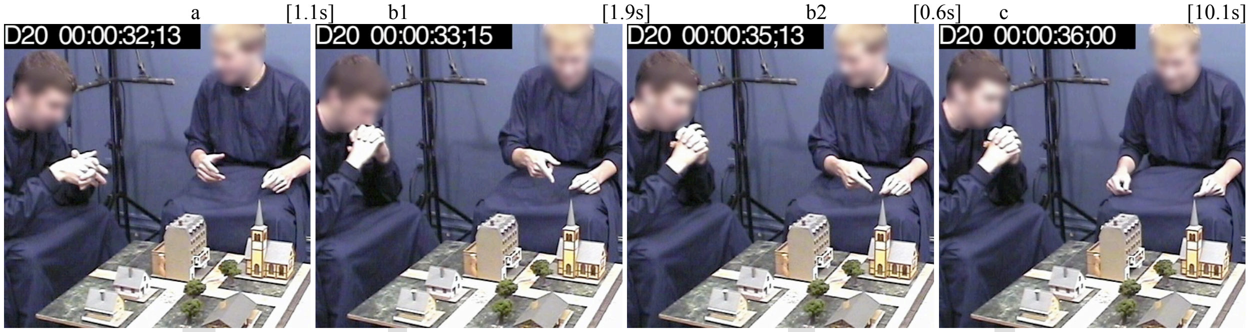
[**A**: No. ·· Thirty-fi[h1]ve. ·· [h2] Thirty-three, Thirty-five. Got it? ··]
[**B**: Oh okay. ·· Right. · It's on that* Alright. Got it. ···]

A: So we go to House Thirty-three. |**B**: Okay. ·· |**A**: And we talk to the people there.

a-b1: *Just before the moment of image* [a], *Participant A's RH forms an index finger deictic toward the church, with no arm extension, as his head and gaze turn to his partner. Following the moment of image* [a], *as he proceeds to the rest of his utterance, he leads into a clearer stroke while extending the arm and looking more closely at the church (see image* [b1]*).*

b1-b2: *His gaze returns almost immediately to his partner's face (this occurs by the end of the word "church"), and he holds the gesture in place (in the same configuration as image* [b2]*) as he finishes the utterance; he continues holding it as he adds an explicit interrogative in speech following a slight pause. His partner, meanwhile, has already begun responding before the interrogative is spoken.*

b2-c: *Participant A retracts his hand to rest as his partner continues responding.*



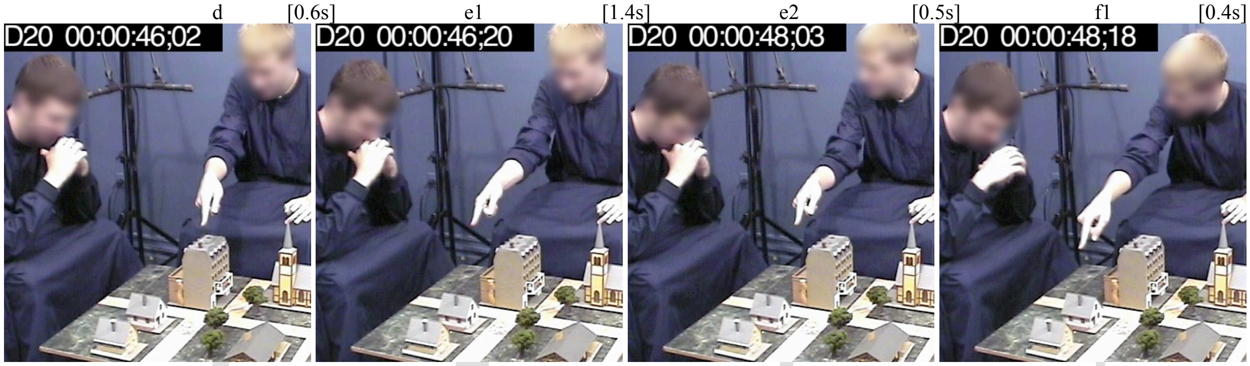
[**A**: and we'll know it because · # the church is right before it. ·· Right?]
[**B**: Okay well I mean yeah I assume the train will stop]
(A's RH) ↗ ^ ----------- ↗ ^ ———————————— ↘ ——————

c-d: *10.1 seconds pass between the moments shown in images* [c] *and* [d], *during which Participant A has continued discussing the plan and has lately begun a RH deictic hold toward the model trees.*

d-e1: *His gaze and RH deictic move to target the most distant house; he jabs at it and holds.*

e1-e2: *Just after the moment shown in image* [e1], *he turns his head and gaze to his partner's face, then jabs at the house again and continues the hold in silence, awaiting a response, which he receives from Participant B just before the moment shown in image* [e2].

e2-f1: *After his partner's response, Participant A extends his arm to more closely pinpoint House 33, while still gazing intently at his partner's face.*



⌈**A**: these t* · these trees and go to House Thirty-three. ··· Thirty-three. ⌉

⌊**B**: Uh-huh. · Thirty- ⌋

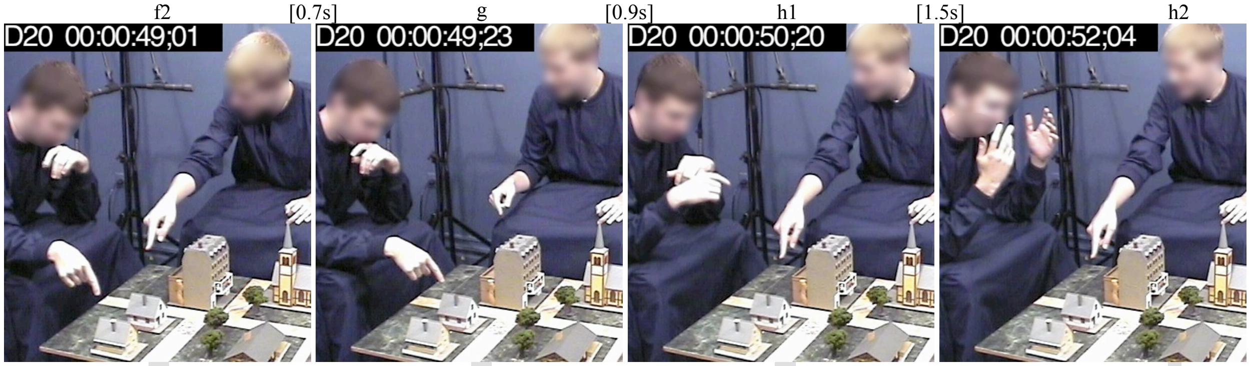
(A's RH) ———————— ↗ ∧ ———————— ∧ ———————————— ↗ ∧ ————————

f1-f2: *As Participant A continues the deictic hold, his partner also pinpoints it with a RH deictic while repeating the house's number in speech.*

f2-g: *Participant A retracts his hand to rest (or nearly so) while repeating the house's number again. Meanwhile, Participant B pinpoints the other house with RH while stating the incorrect number.*

g-h1: *Participant A reaches toward the second house for another deictic hold, while correcting in speech his partner's erroneous assumption of its number, as Participant B retracts his gesture.*

h1-h2: *Participant B briefly repeats the RH deictic toward House 35 (prep. is visible in image* [h1]*), then produces two open-palm gestures with BH near his head (see image* [h2]*), as Participant A continues holding a RH deictic on House 35 while gazing at Participant B's face.*



⌈**A**: Thirty - three. No. ·· Thirty - five. ·· ⌉

⌊**B**: - three, Thirty - two. ·· Oh okay. ·· Right. ⌋

(A's RH) ————— ↘ ˆ _____ ↗ ∧ ——————————————————————————————

(B's RH) ↗ ∧—— ↗ ∧ ——————— ↘ ↗ ∧ ↘ ***(B's BH)*** ↗ ∧ ↘ ↗ ∧ ↘

The hold shown in images [h1] and [h2] is combined with gaze directed at Participant B, thus it is certainly part of a response-seeking utterance. However, it is not a full handoff, because Participant A holds his gesture quite securely as his partner displays corrected understanding. Rather than handing off control of the floor, Participant A intends to continue speaking. This is verified in the full transcript given at the start of the example: following the period shown in the images, his partner attempts to initiate a full utterance, consistent with Participant A's continued gaze. However, Participant A, grinning slightly, appears to wait until precisely that moment to interrupt his partner and once again point at the houses while stating their names, followed by an additional demand for his partner to display understanding. Rather than allow a handoff, he asserts his control (and dominance) in a sort of jokingly punitive manner, consistent with the humorously military framing in which he and his partner will conduct themselves throughout their discourse, with Participant A unambiguously in the dominant role while his partner attempts to be more playful (for example, in Example 4.2 on p. 104, Participant B suggests that they should retain the services of the Crocodile Hunter, and is later chastised for revealing a perverse desire to feast on the wombats they have been charged with capturing).

Maintenance of holds during subordinate responses

The last part of Example 5.3 above demonstrates that among gestures maintained across *planned* turn transitions (which we may wish to distinguish from those that are maintained during interruptions), speakers do not always relinquish control, though they are in a sense still 'handing off' the content of their gesture and intend it to be used by their interlocutor. While seeking a response from an interlocutor by way of gaze coupled with contextual enforcement from a gesture hold, a speaker can simultaneously plan to keep the gesture in play across the interlocutor's utterance, essentially treating the interlocutor's contribution as a controlled insertion or aside within the speaker's ongoing utterance plans. It is an interval in the participants' "joint activity" (Clark 1996) which the gesturer would like to keep rather asymmetrical in terms of who is in charge. Bavelas et al. (2002) have noted a short-duration "gaze window" during which speakers can reassert control after seeking a response via listener-directed gaze, provided they quickly avert their gaze and resume speaking. But by simply maintaining a gesture hold, a speaker can still assert a continuing claim to the floor without the need to avert gaze. Such a move allows for a longer insertion by the interlocutor, with the original speaker's gesture lasting as a 'bridge' across it in a manner rather similar to holds that span *un*requested insertions (e.g., interruptions occurring in the absence of speaker gaze directed at the interrupter, which was the theme of Chapter 4).

Two examples from Chapter 2 follow this formula relatively closely in that the interlocutor's contribution may well have been mobilized by the speaker's gaze, while the speaker still intended to continue speaking eventually. In Example 2.3 (p. 46), the speaker pauses speaking in the middle of producing a 'right-angle' shaped gesture with both hands, while gazing intently at her interlocutor. The interlocutor completes her partner's sentence fragment, suggesting "corner them?" as an appropriate choice fitting both the continuing gesture hold as well as the utterance's preceding speech fragment. The original speaker does not relinquish the hold, and instead goes on to produce additional speech while maintaining it: she repeats the phrase, "corner them," confirming her partner's suggestion, and then performs new gestures while finishing her speech with "and force them to go out the back." Similarly, in Example 2.4 (p. 47), the speaker suspends her utterance after saying "the tarp can't be so big that each person can't… uhm…." During this hiatus she continues to hold a 'hefting' gesture with both arms and gazes at her interlocutor, who suggests "manage it by themselves?" as a sentence completion. The original speaker then agrees and continues speaking. These two examples share several features: in both cases, the original speaker continues gazing at the interlocutor, continues the gesture hold, and resumes speaking upon agreeing with the interlocutor's suggestion. Thus, gesture holds occurring with disfluent or paused speech can become the basis for speaker-requested insertions by the interlocutor, when the speaker directs gaze to mobilize a response. In contrast, the disfluent speaker in Example 2.2 (p. 41) does not direct his gaze at his interlocutor. In attempting to elaborate on his claim that "wombats aren't that big," he repeats the phrase "they're" several times while looking off to the side and waggling a gesture meant to depict an approximate size range. His interlocutor, in turn, waits until he retracts this gesture before responding to his fragmentary but still interpretable utterance.

As mentioned in Chapter 2, Streeck (1993, 1994) discussed a case very similar to Examples 2.3 and 2.4, in which a speaker gestures while searching for a word (the speaker is discussing the experience of being unable to hear her telephone due to having turned it down to a low volume setting). Streeck first argued (1993) that the listener responded to the speaker's word-search disfluency because of the speaker's gaze directed *at her own gesture.* The speaker then briefly becomes the listener, and maintains her gesture while gazing at her interlocutor. This example at first seems troubling given that a *lack* of gaze directed at a listener (e.g., when gaze is directed at one's own gesture instead) should inhibit response, not mobilize it, when in the absence of other response-mobilizing behaviors.[2] However, Streeck's later (1994) paper appears to alter the order of events: the speaker first gazes at her own gesture, and then directs her gaze at her listener just *before* the listener's spoken suggestion, an act which "invites her to join" (Streeck 1994, p. 253). With this correction in the later paper, Streeck's example comes into complete agreement with my own observations of gestures during disfluent episodes, derived from Examples 2.1-2.4: gaze directed at a gesture, and not at the listener, indicates the speaker wishes to retain the floor, whereas gaze directed at the listener can very quickly mobilize a response when the speaker is not actively producing additional speech.[3] This can occur when an utterance is suspended during disfluency (Chapter 2), or when an utterance is complete but gesture and gaze are maintained (the current chapter).

[2] Gullberg and Kita (2009) showed that listeners are affected by speakers directing their gaze at their own gestures, but in a manner independent of the question of listener response: listeners remember significantly more information from such speaker-fixated gestures, when tested after watching a full narration.

[3] Nonetheless, there is likely a great deal of individual and cultural variation affecting whether a listener will feel compelled to insert suggested completions during speaker pauses, including cases where there is no gaze directed at the listener (see Hayashi 2005 for an example of this).

In a sense, the examples just discussed are a blend of the features of 'interruption-bridging', in which an (unsought) contribution is spanned by a persistent gesture whose presence maintains evidence of a speaker's plan to continue, and 'full' handoffs, in which gaze or interrogative speech is directed at the listener (to mobilize response) and the gesture (as well as the claim to continue speaking) is relinquished shortly after the listener takes over speaking duties. The ability to mix and match these criteria results from the fact that gaze direction (to one target or another), speech (or lack thereof), and maintenance (or retraction) of a hold, are activities under independent control: the particular combinations resulting in 'interruption bridging' or full 'handoffs' were frequent enough for me to give them names, but they are still just combinations of more fundamental building blocks.

We might expect that a response-inhibiting, floor-claiming function like 'interruption bridging' and a response-seeking, floor-yielding function like a 'handoff' could not occur simultaneously with a single gesture hold. However, when there are more than two participants, a single gesture can support these seemingly conflicting functions. Such a combination is evident in an example of Sidnell's (2005, pp. 75-78), in which a gesture hold simultaneously persists across an unsought interruption, *and* is simultaneously and independently intended as a 'handoff'. The way the example unfolds is straightforward: one speaker refers to a location in the distance while pointing at it and gazing at the listener he intends to secure a response from, meanwhile that listener and a third participant briefly speak together about something else. This causes a delay in the intended listener's ability to respond to the utterance of the one maintaining the hold, who waits for her to finish her separate speaking task with the third participant. When the intended listener finally responds to the statement that had coincided with the use of the deictic gesture, the gesture can finally be retracted. This speaker's interaction with his intended listener follows the 'handoff' mechanism I have described, meanwhile his gesture hold is also able to span the unwanted hiatus in the listener's ability to engage with him as an interlocutor.

Sidnell's work is of additional interest for its cultural setting (the Caribbean island of Bequia, among speakers of an English-based creole). This is far removed from the setting for my corpus of videos, a University of Chicago laboratory where incredulous undergraduates were asked to ponder an infestation of intelligent wombats. Yet there are potentially universal ways in which the speech, gaze, and gesture of various participants are able to engage with each other and the environment across spans of time, regardless of cultural context or the particularities of the environment or the topics being discussed.

Conclusion

In previous chapters, although I often focused on the potential utility for a listener engaged in comprehension, I also generally showed gesture holds to be part of a speaker's maintenance of expression across spans of time, as well as maintenance of control and a claim to 'speakership' (Schegloff 1984). The current chapter, in contrast, has been somewhat of a departure in that it showed speakers to be capable of also treating holds as *transitive* objects meant to form the basis of a response by the listener. When combined with response-mobilizing cues such as listener-directed gaze (Stivers and Rossano 2010), the content born by the hold is no longer tied to a claim of speakership by the person maintaining the hold, instead it is explicitly meant to enforce a context for a *transfer* of speaking duties: it is part of an expectation of speakership by another. With the retraction of the hold following successful transfer, this shift becomes permanent until another transfer is appropriate. Alternatively, by not 'handing off' the hold and instead maintaining it straight through the listener's contribution, the original speaker can seek a limited response while attempting to maintain the dominant speaking role.

Janet Bavelas and colleagues (Bavelas 1994; Bavelas et al. 1995) have discussed a category of gestures which they dub 'interactive', which play a role in turn transitions, and which can be contrasted with 'topical' gestures related to the content of speech.[4] But all gestures spanning turn transitions are unavoidably interactive, whether or not they also support the content of the speakers' utterances. In my examples, listener-directed gaze during an extended hold appears to create a combination that can play a part in fulfilling some of the same interactive, turn-transitioning functions as those described by Bavelas and colleagues, but which also involve unequivocally 'topical' gestures. By extending into the silence beyond a speaker's fluently or disfluently ended speech, these 'handoff' holds continue to provide topical information while being presented for use in the listener's response. Furthermore, rather than require a separate interactive gesture for yielding a turn, the retraction of a previously begun hold, right after a successful transfer of speaking duties, can also be interpreted as a purposeful yielding of control. Allowing for 'topical' gestures to fulfill simultaneous 'interactive' functions is consistent with Bavelas' own positive stance regarding the ability of gestures to fulfill more than one function at once (Bavelas 1994, p. 204).

[4] Included in their subcategories for interactive gestures (Bavelas 1994, p. 213; Bavelas et al. 1995, p. 397) are functions such as "giving turn" (which "hands over" the speaking turn), "seeking following" (which gauges a listener's understanding), and "seeking help" (when a speaker is having trouble finding a word). The authors state that these interactive gestures "serve several functions necessary for dialogue" but that they "provide no information about the topic at hand."

6 Concluding remarks

In the chapters above I have moved from a treatment of gesture holds occurring as part of a speaker's own utterance production processes (Chapters 2 and 3), to holds occurring because of competitive contributions by an interlocutor (Chapter 4), and finally to holds occurring at the bridge between the two, when a speaker seeks to link a completed utterance to a new contribution by an interlocutor (Chapter 5).

In order to offer a new perspective on the potential functions of these spans of 'embodied stasis', one that goes beyond McNeill's (1992, 2005) well-known production-oriented theories, I have emphasized the potential benefits for the *listener,* in the chapters illustrating gestural behavior during a speaker's own disfluent utterances (Chapters 2 and 3). In such cases, the asymmetry or complementarity between spoken and gestured content can temporarily become large, due to a hiatus in the speech stream co-occurring with continued representation in gesture. Taking speech alone, this would seem to coincide with an especially large asymmetry between a speaker's awareness of in-progress utterance formulation and delivery, and a listener's ability to make timely interpretations from fragmentary evidence. With continuous observation of the 'whole' speaker, however, I showed that the listener has access to a great deal of evidence from gesture, made available for observation across significant spans of time (perhaps lengthened because of the hiatus in speech). Though not always overtly synchronized with speech in the manner of a stroke lining up with a prosodic peak, this kind of gestural evidence is far from asynchronous with the utterance: it provides imagery for what a speaker *would* be speaking if not for the hiatus, especially in the case of 'prospective' gestural cues revealing aspects of the fully fluent speech-gesture composite to come. Some information available in gesture can be composed smoothly with initial elements of a constituent that have emerged in speech just before a hiatus, such as a determiner or other bound form spoken as a preliminary commitment to utterance. Gestural representation can *fulfill* some of this commitment with content presaging the identity of the syntactically required speech that must follow the hiatus (Chapter 3).

In the case of 'retrospective' gesture holds following successfully completed pulses of utterance, the effect on the listener is not so much one of revealing what a speaker *would* be saying if not for the pause, but rather an 'interactive' (Bavelas 1994) effect of maintaining a speaker's claim of 'speakership' (Schegloff 1984) coupled with the fact that such gestures are also forward-contextualizing: they 'topically' indicate that the imagery of the preceding utterance is still valid during the pause and will be the starting point for the subsequent resumption. This is significantly different from McNeill's (1992) and Kita's (1990, 1993; Kita et al. 1998) conception of the role for the post-stroke hold, because it reaches beyond the bounds of the preceding utterance, in which its prolongation is said to have come about as an extension of the stroke around co-emergent speech, rather than future formulations. Instead, extended claims to speakership, as well as extended thematic contextualization or 'bracketing', can be displayed indefinitely through the prolonged maintenance of such holds, with one limb remaining in place while new speech-gesture moves are performed with another, often including overlapping holds that allow for the hands to trade off (recall Chapter 2 on gestural 'buoys', Examples 2.9 and 2.10). In Chapter 5, however, I showed that these forward-contextualizing, 'retrospective' holds lasting beyond the boundaries of utterances can also be employed for the explicit *transfer* of speaking duties, when coupled with response-mobilizing behaviors such as listener-directed gaze or interrogative speech (Stivers and Rossano 2010). Thus while the hold may maintain a claim to speakership, it is a claim which can be imposed on the listener instead, a hot potato of responsibility to be 'handed off' along with its thematic material, with the hold relinquished as soon as this transfer is achieved.

Chapter 4, meanwhile, showed that holds are not just concomitants to 'personal' disfluencies and pauses in speech, but also to disruptions instigated by interlocutors. In either case, the form is the same: an evolving process of gestural movement suspends and instead enters an interval of stasis. The interactional properties, however, are quite different, because in the case of interruptions the interlocutor is not passively striving to understand a fragmentary utterance, and is instead the root cause of that fragmentation. Therefore, I focused on the possibility of a speaker's persistent gestural anchor to store, for future use, the state of the utterance at the point it was disrupted. A discussion of current knowledge regarding the properties and constraints of human attention and memory led me to conclude that gesture holds, available as relatively stable (though temporary) 'artifacts', should be capable of alleviating some of the need for speakers and listeners to internally rehearse the information being built up across and within utterances. In the case of interruptions that could completely distract the interactants from their existing conversational projects, such gestural artifacts would also automatically 'bridge' the disruption and remain available afterward as *recall cues* (Nelson and Goodmon 2003).

Throughout, the issue of the timing of the hold relative to ongoing speech has been of paramount importance, because it determines the manner in which the hold supports temporal cohesion across the discourse. In the case of 'prospective' holds occurring near the start of utterances, the gesture binds *present* to *future* by presaging the spoken content that has not yet arrived in speech. When gestures 'retrospectively' maintain a configuration already co-expressive with a previous full utterance, they bind *present* to *past.* These relationships are not mutually exclusive: during the span of any hold, but most relevantly during those spanning multiple utterances, the hold can bind *past* to *present* to *future* in a framing or bracketing manner that unifies all of the encompassed discourse into a thematic whole. Through a single core function, which is simply to hold or *maintain* (McCullough 2005), holds are both completely dependent on our physical and temporal reality, while also providing a means for us to overcome some of its constraints.

Directions for future research

In this dissertation I have only scratched the surface of what may be an enormous array of functions relating to our engagement with temporarily static configurations such as gesture holds. The following are just a few avenues for future work, among many others.

First, I should emphasize that in nearly all cases, I have discussed representational and pointing gestures with clearly defined, object referents. I did not yet explore how these gestural artifacts may also, perhaps simultaneously, index other kinds of less concretely defined meaning, including the social structures the objects are bound within (see Hanks 2005). Because of this emphasis on concrete single referents, I also provided comparatively little discussion of 'bracketing' or framing by gesture holds, which can occur not only via concrete reference serving to metonymically index a larger thematic scope (as with long-lasting gestural 'buoys'), but also via gestures serving "pragmatic" functions, as have been frequently discussed by Kendon (2004, inter alia, regarding gesture "families" such as open-handed gestures with the palm tilted up, and those with the palm tilted down).

Regarding the experimental evidence discussed in Chapter 3, such as the eye-tracking studies showing that listeners respond to incremental speech (including disfluencies) by narrowing the scope of likely reference (Arnold et al. 2003, 2004), an obvious step would be to conduct similar experiments which incorporate gestures integrated naturally into the stimulus materials, to see if they indeed help to further narrow the scope of likely reference in the manner I have suggested. That chapter also offered intriguing hints that a more linguistic focus may be possible for studying the compositionality of speech and gesture bearing complementary content: the syntactic status of individual words (such as the requirement that an object slot be filled with additional speech) may affect their compositionality with *simultaneous* gestures that can, perhaps, temporarily take the place of linguistic material that has not yet arrived in the stream of speech. This was shown, for example, in Examples 3.1 and 3.2, in which a determiner (*those* and *the,* respectively) was synchronized with an iconic or deictic gesture that created strong expectations regarding syntactically required speech—speech which did not arrive until after a significant delay.

A third area for future research is the integration of gestural artifacts such as those I have described with built structures and artifacts external to our bodies, that are in fact far more stable across time. This is the realm of 'distributed cognition' (Hutchins 1995, 2006) as well as socially sedimented cultural practice (Hanks 2005, inter alia), also incorporating 'environmentally coupled' gestures (Goodwin 2007). Any investigations into the capacity of gestural artifacts to act as 'recall cues' should also keep in mind the role of built spaces, recruited tools, and other external creations in the service of maintaining cohesion across disruptive events and digressions in conversation.

References

Altmann, G. T. M. and Y. Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* 73: 247-264.

Arnold, Jennifer E., Maria Fagnano, and Michael K. Tanenhaus. 2003. Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research* 32 (1): 25-36.

Arnold, Jennifer E., Michael K. Tanenhaus, Rebecca J. Altmann, and Maria Fagnano. 2004. The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science* 15 (9): 578-582.

Austin, John L. 1962. *How to Do Things With Words.* Cambridge, MA: Harvard University Press.

Baddeley, Alan D. 1986. *Working Memory.* Oxford: Clarendon Press.

Baddeley, Alan D. 2001. Is working memory still working? *American Psychologist* 56: 849-864.

Baddeley, Alan D. and Robert H. Logie. 1999. Working memory: The multiple-component model. In *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control,* edited by Akira Miyake and Priti Shah. Cambridge: Cambridge University Press, pp. 28-61.

Bartlett, F. C. 1932. *Remembering.* Oxford: Oxford University Press.

Bavelas, Janet Beavin. 1994. Gestures as part of speech: Methodological implications. *Research on Language and Social Interaction* 27 (3): 201-221.

Bavelas, Janet Beavin, Nicole Chovil, Linda Coates, and Lori Roe. 1995. Gestures specialized for dialogue. *Personality and Social Psychology Bulletin* 21 (4): 394-405.

Bavelas, Janet Beavin, Linda Coates, and Trudy Johnson. 2002. Listener responses as a collaborative process: The role of gaze. *Journal of Communication* 52 (3): 566-580.

Beattie, Geoffrey and Heather Shovelton. 1999a. Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica* 123 (1/2): 1-30.

Beattie, Geoffrey and Heather Shovelton. 1999b. Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology* 18 (4): 438-462.

Beattie, Geoffrey and Heather Shovelton. 2001. An experimental investigation of the role of different types of iconic gesture in communication: A semantic feature approach. *Gesture* 1 (2): 129-149.

Butterworth, Brian and Geoffrey Beattie. 1978. Gesture and silence as indicators of planning in speech. In *Recent Advances in the Psychology of Language: Formal and Experimental Approaches, Part 2,* edited by Robin N. Campbell and Philip T. Smith. New York: Plenum Press, pp. 347-360.

Butterworth, Brian and Uri Hadar. 1989. Gesture, speech, and computational stages: A reply to McNeill. *Psychological Review* 96 (1): 168-74.

Chafe, Wallace. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing.* Chicago: The University of Chicago Press.

Chambers, Craig G., Michael K. Tanenhaus, Kathleen M. Eberhard, Hana Filip, and Greg N. Carlson. 2002. Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language* 47: 30-49.

Clark, Herbert H. 1996. *Using Language.* Cambridge: Cambridge University Press.

Clark, Herbert H. and Thomas Wasow. 1998. Repeating words in spontaneous speech. *Cognitive Psychology* 37: 201-242.

Condon, W. S. and W. D. Ogston. 1966. Sound film analysis of normal and pathological behavior patterns. *Journal of Nervous and Mental Disease* 143 (4): 338-347.

Condon, W. S. and W. D. Ogston. 1967. A segmentation of behavior. *Journal of Psychiatric Research* 5 (3): 221-235.

Cowan, Nelson. 1988. Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin* 104 (2): 163-191.

Cowan, Nelson. 1999. An embedded-processes model of working memory. In *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control,* edited by Akira Miyake and Priti Shah. Cambridge: Cambridge University Press, pp. 62-101.

Dancygier, Barbara and Eve Sweetser. 2005. *Mental Spaces in Grammar: Conditional Constructions.* Cambridge: Cambridge University Press.

De Ruiter, Jan-Peter. 1998. Gesture and speech production. Doctoral dissertation, Catholic University of Nijmegen. *MPI Series in Psycholinguistics, 6.* Max Planck Institute for Psycholinguistics, Nijmegen.

De Ruiter, Jan-Peter. 2000. The production of gesture and speech. In *Language and Gesture,* edited by David McNeill. Cambridge: Cambridge University Press, pp. 284-311.

De Ruiter, Jan-Peter. 2003. The function of hand gesture in spoken conversation. In *Manus Loquens: Medium der Geste, Gesten der Medien,* edited by Matthias Bickenbach, Annina Klappert, and Hedwig Pompe. Cologne: DuMont, pp. 338-347.

De Ruiter, Jan-Peter, Holger Mitterer, and N. J. Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language* 82 (3): 515-535.

Dell, Cecily. 1970. *A Primer for Movement Description: Using Effort-Shape and Supplementary Concepts.* New York: Dance Notation Bureau, Inc.

Downing, Pamela. 1977. On the creation and use of English compound nouns. *Language* 53 (4): 810-842.

Du Bois, John W., Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino. 1993. Outline of discourse transcription. In *Talking Data: Transcription and Coding in Discourse Research,* edited by Jane A. Edwards and Martin D. Lampert. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., pp. 45-90.

Duncan, Starkey, Jr. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology* 23 (2): 283-292.

Duncan, Starkey, Jr. 1973. Toward a grammar for dyadic conversation. *Semiotica* 9 (1): 29-46.

Duncan, Susan D. 1996. Grammatical form and 'thinking-for-speaking' in Mandarin Chinese and English: An analysis based on speech-accompanying gestures. Ph.D. dissertation, University of Chicago.

Duncan, Susan D. 2008. Gestural imagery and cohesion in normal and impaired discourse. In *Embodied Communication in Humans and Machines,* edited by Ipke Wachsmuth, Manuela Lenzen, and Günther Knoblich. New York: Oxford University Press, pp. 305-328.

Eberhard, Kathleen M., Michael J. Spivey-Knowlton, Julie C. Sedivy, and Michael K. Tanenhaus. 1995. Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research* 24 (6): 409-436.

Ekman, Paul and Wallace V. Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica* 1: 49-98.

Enfield, N. J. 2004. On linear segmentation and combinatorics in co-speech gesture: A symmetry-dominance construction in Lao fish trap descriptions. *Semiotica* 149: 57-123.

Enfield, N. J. 2005. The body as a cognitive artifact in kinship representations: Hand gesture diagrams by speakers of Lao. *Current Anthropology* 46 (1): 51-81.

Enfield, N. J. 2009. *The Anatomy of Meaning: Speech, Gesture, and Composite Utterances.* Cambridge: Cambridge University Press.

Enfield, N. J. and Stephen C. Levinson. 2006. Introduction: Human sociality as a new interdisciplinary field. In *Roots of Human Sociality: Culture, Cognition and Interaction,* edited by N. J. Enfield and Stephen C. Levinson. Oxford and New York: Berg, pp. 1-35.

Fauconnier, Gilles. 1994. *Mental Spaces: Aspects of Meaning Construction in Natural Language.* Cambridge: Cambridge University Press.

Fauconnier, Gilles. 1997. *Mappings in Thought and Language.* Cambridge: Cambridge University Press.

Fillmore, Charles J. 1976. Frame semantics and the nature of language. In *Origins and Evolution of Language and Speech,* edited by Stevan R. Harnad, Horst D. Steklis, and Jane Beckman Lancaster. New York Academy of Sciences, Vol. 280, pp. 20-32.

Fillmore, Charles J. 1985. Frames and the semantics of understanding. *Quaderni di Semantica* 6 (2): 222-254.

Furuyama, Nobuhiro. 2000. Gestural interaction between the instructor and the learner in *origami* instruction. In *Language and Gesture,* edited by David McNeill. Cambridge: Cambridge University Press, pp. 99-117.

Goffman, Erving. 1974. *Frame Analysis.* New York: Harper and Row.

Goldin-Meadow, Susan. 2003. *Hearing Gesture: How Our Hands Help Us Think.* Cambridge, MA: Belknap Press of Harvard University Press.

Goldin-Meadow, Susan. 2006. Meeting other minds through gesture: How children use their hands to reinvent language and distribute cognition. In *Roots of Human Sociality: Culture, Cognition and Interaction,* edited by N. J. Enfield and Stephen C. Levinson. Oxford and New York: Berg, pp. 353-374.

Goldin-Meadow, Susan, Howard Nusbaum, Spencer D. Kelly, and Susan Wagner. 2001. Explaining math: Gesturing lightens the load. *Psychological Science* 12 (6): 516-522.

Goodwin, Charles. 1986. Gestures as a resource for the organization of mutual orientation. *Semiotica* 62 (1/2): 29-49.

Goodwin, Charles. 2000. Action and embodiment within situated human interaction. *Journal of Pragmatics* 32 (10): 1489-1522.

Goodwin, Charles. 2006. Human sociality as mutual orientation in a rich interactive environment: Multimodal utterances and pointing in aphasia. In *Roots of Human Sociality: Culture, Cognition and Interaction,* edited by N. J. Enfield and Stephen C. Levinson. Oxford and New York: Berg, pp. 97-125.

Goodwin, Charles. 2007. Environmentally coupled gestures. In *Gesture and the Dynamic Dimension of Language,* edited by Susan D. Duncan, Justine Cassell, and Elena T. Levy. Amsterdam: John Benjamins, pp. 195-212.

Gullberg, Marianne and Kenneth Holmqvist. 1999. Keeping an eye on gestures: Visual perception of gestures in face-to-face communication. *Pragmatics & Cognition* 7 (1): 35-63.

Gullberg, Marianne and Kenneth Holmqvist. 2006. What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video. *Pragmatics & Cognition* 14 (1): 53-82.

Gullberg, Marianne and Sotaro Kita. 2009. Attention to speech-accompanying gestures: Eye movements and information uptake. *Journal of Nonverbal Behavior* 33: 251-277.

Gumperz, John J. 1982. *Discourse Strategies.* Cambridge: Cambridge University Press.

Hanks, William F. 1983. Deixis and the organization of interactive context in Yucatec Maya. Ph.D. dissertation, University of Chicago.

Hanks, William F. 2005. Explorations in the deictic field. *Current Anthropology* 46 (2): 191-220.

Hayashi, Makoto. 2005. Joint turn construction through language and the body: Notes on embodiment in coordinated participation in situated activities. *Semiotica* 156 (1/4): 21-53.

Hockett, Charles F. 1958. *A Course in Modern Linguistics.* New York: Macmillan.

Hutchins, Edwin. 1995. *Cognition in the Wild.* Cambridge, MA: The MIT Press.

Hutchins, Edwin. 1996. Cognitive artifacts. In *The MIT Encyclopedia of the Cognitive Sciences,* edited by Robert A. Wilson and Frank C. Keil. Cambridge, MA: The MIT Press, pp. 126-128.

Hutchins, Edwin. 2006. The distributed cognition perspective on human interaction. In *Roots of Human Sociality: Culture, Cognition and Interaction,* edited by N. J. Enfield and Stephen C. Levinson. Oxford and New York: Berg, pp. 375-398.

Jefferson, Gail. 1989. Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In *Conversation: An Interdisciplinary Perspective,* edited by Derek Roger and Peter Bull. Clevedon, UK and Philadelphia: Multilingual Matters Ltd, pp. 166-196.

Kelly, Spencer D., Corinne Kravitz, and Michael Hopkins. 2004. Neural correlates of bimodal speech and gesture comprehension. *Brain and Language* 89: 253-260.

Kendon, Adam. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica* 26: 22-63.

Kendon, Adam. 1972. Some relationships between body motion and speech: An analysis of an example. In *Studies in Dyadic Communication,* edited by Aron Wolfe Siegman and Benjamin Pope. New York: Pergamon Press, pp. 177-210.

Kendon, Adam. 1980. Gesticulation and speech: Two aspects of the process of utterance. In *The Relationship of Verbal and Nonverbal Communication,* edited by Mary Ritchie Key. The Hague: Mouton, pp. 207-227.

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance.* Cambridge: Cambridge University Press.

Kita, Sotaro. 1990. The temporal relationship between gesture and speech: A study of Japanese-English bilinguals. Master's thesis, University of Chicago.

Kita, Sotaro. 1993. Language and thought interface: A study of spontaneous gestures and Japanese mimetics. Ph.D. dissertation, University of Chicago.

Kita, Sotaro. 2000. How representational gestures help speaking. In *Language and Gesture,* edited by David McNeill. Cambridge: Cambridge University Press, pp. 162-185.

Kita, Sotaro, Ingeborg van Gijn, and Harry van der Hulst. 1998. Movement phases in signs and co-speech gestures, and their transcription by human coders. In *Gesture and Sign Language in Human-Computer Interaction,* edited by Ipke Wachsmuth and Martin Fröhlich. Proceedings of the International Gesture Workshop, Bielefeld, Germany, September 17-19, 1997. Berlin: Springer-Verlag, pp. 23-35.

Krauss, Robert M., Yihsiu Chen, and Rebecca F. Gottesman. 2000. Lexical gestures and lexical access: A process model. In *Language and Gesture,* edited by David McNeill. Cambridge: Cambridge University Press, pp. 261-283.

Krauss, Robert M., Palmer Morrel-Samuels, and Christina Colasante. 1991. Do conversational hand gestures communicate? *Journal of Personality and Social Psychology* 61 (5): 743-754.

Lakoff, George. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind.* Chicago: The University of Chicago Press.

Levelt, Willem J. M. 1983. Monitoring and self-repair in speech. *Cognition* 14: 41-104.

Levelt, Willem J. M. 1989. *Speaking: From Intention to Articulation.* Cambridge, MA: The MIT Press.

Levinson, Stephen C. 1978. Activity types and language. *Pragmatics Microfiche* 3: 3-3, D1-G5.

Levinson, Stephen C. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature.* Cambridge, MA: The MIT Press.

Levinson, Stephen C. 2006. On the human "interaction engine". In *Roots of Human Sociality: Culture, Cognition and Interaction,* edited by N. J. Enfield and Stephen C. Levinson. Oxford and New York: Berg, pp. 39-69.

Liddell, Scott K. 2003. *Grammar, Gesture, and Meaning in American Sign Language.* Cambridge: Cambridge University Press.

Linell, Per. 1998. *Approaching Dialogue: Talk, Interaction and Contexts in Dialogical Perspectives.* Amsterdam: John Benjamins.

Loehr, Daniel. 2007. Aspects of rhythm in gesture and speech. *Gesture* 7 (2): 179-214.

Matsumoto, Yoshiko. 1997. *Noun-Modifying Constructions in Japanese: A Frame-Semantic Approach.* Amsterdam: John Benjamins.

Mayberry, Rachel I. and Joselynne Jaques. 2000. Gesture production during stuttered speech: Insights into the nature of gesture-speech integration. In *Language and Gesture,* edited by David McNeill. Cambridge: Cambridge University Press, pp. 199-214.

McCullough, Karl-Erik. 2005. Using gestures during speaking: Self-generating indexical fields. Ph.D. dissertation, University of Chicago.

McNeill, David. 1985. So you think gestures are nonverbal? *Psychological Review* 92 (3): 350-371.

McNeill, David. 1989. A straight path—to where? Reply to Butterworth and Hadar. *Psychological Review* 96 (1): 175-179.

McNeill, David. 1992. *Hand and Mind: What Gestures Reveal about Thought.* Chicago: The University of Chicago Press.

McNeill, David. 2000. Catchments and contexts: Non-modular factors in speech and gesture production. In *Language and Gesture,* edited by David McNeill. Cambridge: Cambridge University Press, pp. 312-328.

McNeill, David. 2005. *Gesture and Thought.* Chicago: The University of Chicago Press.

McNeill, David, Elena T. Levy, and Laura L. Pedelty. 1990. Speech and gesture. In *Cerebral Control of Speech and Limb Movements (Advances in Psychology, 70),* edited by Geoffrey R. Hammond. Amsterdam: Elsevier/North-Holland, pp. 203-256.

Miller, George A., Eugene Galanter, and Karl H. Pribram. 1960. *Plans and the structure of behavior.* New York: Holt.

Morrel-Samuels, Palmer and Robert M. Krauss. 1992. Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18 (3): 615-622.

Morsella, Ezequiel and Robert M. Krauss. 2004. The role of gestures in spatial working memory and speech. *American Journal of Psychology* 117 (3): 411-424.

Narayan, Shweta. 2010. *Maybe what it means is he actually got the spot*: Physical and cognitive viewpoint in a gesture study. Ms., University of California, Berkeley.

Nelson, Douglas L. and Leilani B. Goodmon. 2003. Disrupting attention: The need for retrieval cues in working memory theories. *Memory & Cognition* 31 (1): 65-76.

Norman, Donald A. 1991. Cognitive artifacts. In *Designing Interaction: Psychology at the Human-Computer Interface,* edited by J. M. Carroll. Cambridge: Cambridge University Press, pp. 17-38.

Ochs, Elinor. 1979. Transcription as theory. In *Developmental Pragmatics,* edited by Elinor Ochs and Bambi B. Schieffelin. New York: Academic Press, pp. 43-72.

Özyürek, Aslı, Roel M. Willems, Sotaro Kita, and Peter Hagoort. 2007. On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience* 19 (4): 605-616.

Park-Doob, Mischa. 2007a. Gestural effects on working memory and attention. Paper presented at the 3rd Conference of the International Society for Gesture Studies, Northwestern University, Evanston, IL, June 18-21.

Park-Doob, Mischa. 2007b. 'Hold that thought': Effects of held gestures on referent accessibility. Paper presented at the Institute of Cognitive and Brain Sciences faculty retreat, University of California, Berkeley, Dec. 7.

Parrill, Fey. 2009. Dual viewpoint gestures. *Gesture* 9 (3): 271-289.

Pashler, Harold E. 1997. *The Psychology of Attention.* Cambridge, MA: The MIT Press.

Peters, Michael. 1990. Interaction of vocal and manual movements. In *Cerebral Control of Speech and Limb Movements (Advances in Psychology, 70),* edited by Geoffrey R. Hammond. Amsterdam: Elsevier/North-Holland, pp. 535-574.

Reddy, Michael J. 1979. The conduit metaphor: A case of frame conflict in our language about language. In *Metaphor and Thought,* edited by A. Ortony. Cambridge: Cambridge University Press, pp. 284-297.

Sacks, Harvey. 1989. Lectures 1964-1965, edited by Gail Jefferson with an Introduction / Memoir by E. A. Schegloff. *Human Studies* 12 (3-4): 183-404.

Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50 (4): 696-735.

Schegloff, Emanuel A. 1968. Sequencing in conversational openings. *American Anthropologist* 70 (6): 1075-1095.

Schegloff, Emanuel A. 1984. On some gestures' relation to talk. *In Structures of Social Action: Studies in Conversation Analysis,* edited by J. Maxwell Atkinson and John Heritage. Cambridge: Cambridge University Press, pp. 266-296.

Schutz, Alfred. 1967 [1932]. *The Phenomenology of the Social World. Northwestern University Studies in Phenomenology and Existential Philosophy.* Evanston, IL: Northwestern University Press.

Schutz, Alfred. 1973. *Collected Papers, Vol. 1: The Problem of Social Reality,* edited by Maurice Natanson. The Hague: Martinus Nijhoff.

Seyfeddinipur, Mandana. 2006. Disfluency: Interrupting speech and gesture. Doctoral dissertation, Radboud University of Nijmegen. *MPI Series in Psycholinguistics, 39.* Max Planck Institute for Psycholinguistics, Nijmegen.

Seyfeddinipur, Mandana and Sotaro Kita. 2005. Gestures and self-monitoring in speech production. In *Proceedings of the 27th Annual Meeting of the Berkeley Linguistics Society, Feb. 16-18, 2001, General Session and Parasession on Gesture and Language,* edited by Charles Chang, Michael J. Houser, Yuni Kim, David Mortensen, Mischa Park-Doob, and Maziar Toosarvandani. Berkeley, CA: Berkeley Linguistics Society, pp. 457-464.

Sidnell, Jack. 2005. Gesture in the pursuit and display of recognition: A Caribbean case study. *Semiotica* 156 (1/4): 55-87.

Sidnell, Jack. 2006. Coordinating gesture, talk, and gaze in reenactments. *Research on Language and Social Interaction* 39 (4): 377-409.

Slobin, Dan I. 1987. Thinking for speaking. In *Proceedings of the 13th Annual Meeting of the Berkeley Linguistics Society, Feb. 14-16, 1987, General Session and Parasession on Grammar and Cognition,* edited by Jon Aske, Natasha Beery, Laura Michaelis, and Hana Filip. Berkeley, CA: Berkeley Linguistics Society, pp. 435-444.

Slobin, Dan I. 1996. From "thought and language" to "thinking for speaking". In *Rethinking Linguistic Relativity,* edited by John J. Gumperz and Stephen C. Levinson. Cambridge: Cambridge University Press, pp. 70-96.

Smith, Nathaniel. 2003. Gestures and beyond. B.A. thesis, University of California, Berkeley.

Stivers, Tanya and Federico Rossano. 2010. Mobilizing response. *Research on Language and Social Interaction* 43 (1): 3-31.

Streeck, Jürgen. 1993. Gesture as communication I: Its coordination with gaze and speech. *Communication Monographs* 60: 275-299.

Streeck, Jürgen. 1994. Gesture as communication II: The audience as co-author. *Research on Language and Social Interaction* 27 (3): 239-267.

Sweetser, Eve and Gilles Fauconnier. 1996. Cognitive links and domains: Basic aspects of Mental Space Theory. In *Spaces, Worlds, and Grammar,* edited by Gilles Fauconnier and Eve Sweetser. Chicago: The University of Chicago Press, pp. 1-28.

Tuite, Kevin. 1993. The production of gesture. *Semiotica* 93 (1/2): 83-105.

Tulving, Endel and Donald M. Thomson. 1973. Encoding specificity and retrieval processes in episodic memory. *Psychological Review* 80 (5): 352-373.

Vygotsky, Lev S. 1986 [1934]. *Thought and Language.* Edited and translated by E. Hanfmann and G. Vakar [1962], revised and edited by A. Kozulin [1986]. Cambridge, MA: The MIT Press.

Welford, A. T., ed. 1980. *Reaction Times.* New York: Academic Press.

Wesp, Richard, Jennifer Hesse, Donna Keutmann, and Karen Wheaton. 2001. Gestures maintain spatial imagery. *American Journal of Psychology* 114 (4): 591-600.

Willems, Roel M., Aslı Özyürek, and Peter Hagoort. 2007. When language meets action: The neural integration of gesture and speech. *Cerebral Cortex* 17 (10): 2322-2333.

Woodall, W. Gill and Joseph P. Folger. 1985. Nonverbal cue context and episodic memory: On the availability and endurance of nonverbal behaviors as retrieval cues. *Communication Monographs* 52: 319-333.

Wu, Ying Choon and Seana Coulson. 2007. Iconic gestures prime related concepts: An ERP study. *Psychonomic Bulletin & Review* 14 (1): 57-63.