

Ask Not What the Parts of Speech are, Ask What Speech is Part of:  
Toward a Direct Realism of Communication

Gautam K. Vallabha,  
Center for the Neural Basis of Cognition  
Carnegie Mellon University

Bharath Vallabha,  
Department of Philosophy  
Harvard University

This manuscript has been submitted to *Ecological Psychology*.  
Please address questions or comments to:

Gautam K .Vallabha  
4400 Fifth Avenue  
Mellon Institute Room 110  
Pittsburgh, PA 15213, U.S.A.  
Email: vallabha@cnbc.cmu.edu

## Abstract

Speech perception and language in general have been seen as challenges for the ecological approach to cognition. Drawing on insights from ordinary-language philosophy and phenomenology, it is argued that the ecological approach is in fact essential for a full account of linguistic phenomena. The argument critiques four conventional assumptions: (1) speech is the consequence of a thought, (2) listeners recover the speaker's thought from the speech, (3) an utterance is composed of meaningless phonological units, and (4) a word gains meaning through association. These assumptions are shown to be inadequate and a new theoretical stance is advocated, based on the tenets that speaking is a form of thinking, that listening is a public participation in the speaker's thought, and that meaning is defined by use. This stance fits naturally with direct realism, connects with theories of speech that emphasize systematic fine phonetic detail, and clarifies the relation between speech perception and the perception of articulatory events.

## Table of Contents

Direct Realism and Language .....	5
Critique of Traditional Assumptions.....	6
Speaking is Not Separate From Thinking .....	7
Speech is Not Preceded By a “Decision to Speak” .....	8
Speech is Not the Product of a Thought.....	9
Speech is a Form of Thought .....	10
The Speaker's Thoughts are Shaped by the Listener and the Situation.....	11
Listening is Not A Prelude to Understanding .....	13
The Listener Does Not Reconstruct the Speaker's Thought.....	13
The Listener is Not Fundamentally Uncertain About the Speaker's Thought.....	15
Listening is an Active Relation of the Listener to the Speaker .....	17
The Utterance is Not Composed of Phonological Units .....	18
A Word is Defined by Function, Not By Acoustic Structure.....	18
Linguistic and Indexical Information Cannot be Separated.....	20
The Meaning of an Utterance is Shaped By its Use.....	21
Synthesis.....	25
Relation to Direct Realism .....	27
Experimental Approaches .....	29
Reflections.....	31

Ask Not What the Parts of Speech are, Ask What Speech is Part of:  
Toward a Direct Realism of Communicatio

The key premise of direct realism is that an organism's interaction with the world is unmediated (Mace, 1977; Heft, 2001). One consequence of this premise, which has much in common with embodied approaches to cognition (e.g., Varela, Thompson & Rosch, 1991, O'Regan & Noë, 2001), is the inseparability of perception and action: to perceive an entity is not to register it in a covert mental domain, but rather to interact in a manner characteristic to that entity. This approach has been elaborated in detail for a person exploring an inanimate environment, for example, by hefting a stick, seeing a tree, or avoiding collision. It has also been applied to some animate interactions, such as rhythmic coordination between individuals (e.g., Richardson, Marsh, and Schmidt, 2005). However, there is little consensus for how to apply it to social skills such as perceiving fear, threat or friendliness, or social interactions such as greeting, concurring or insulting. The example *par excellence* for social situations is linguistic interaction, and in particular spoken conversational interaction. From a direct realist or embodied perspective, what does it mean to perceive and understand spoken language?

The primary goal of this paper is to propose an answer to the above question. While this issue has been tackled before (e.g., Verbrugge, 1985; Reed, 1996), these earlier attempts started with the premises of direct realism and tried to apply them to linguistic issues. This approach has two problems. First, the argument is persuasive only to those who already accept the premises of direct realism. Second, the approach tacitly accepts some traditional assumptions of language (e.g., that utterances are composed of phonological units) that are conceptually problematic. We suggest that it is more fruitful to engage directly with the traditional definitions, critique the problems therein, and show how a full consideration of linguistic phenomena leads naturally to a direct realist approach. In going about this, we shall also consider two positions articulated by others (the direct perception of articulatory events, and the direct perception of communicative events) and attempt to clarify the relation between them.

A second goal of the paper is to bring out the utility of phenomenology and ordinary-language philosophy to ecological psychology. As noted above, some of the traditional assumptions underlying language are conceptually problematic. Uncovering and critiquing these assumptions is a necessary prelude to *any* theory of speech and language (direct realist or

otherwise). However, as the philosopher J.L. Austin noted, “there is no simple way of doing this - partly because ... there is no simple ‘argument’. It is a matter of unpicking, one by one, a mass of seductive (mainly verbal) fallacies, of exposing a wide variety of concealed motives” (Austin, 1962, p.4). In order to “unpick” these fallacies, one needs to draw attention to the subtleties of our own lived experience and of how we actually use language. The former approach is phenomenology (Heidegger, 1927; Sartre, 1943; Merleau-Ponty, 1962; for an introduction, see Dreyfus, 1991, 1992) and the latter is ordinary-language analysis (Ryle, 1949; Wittgenstein, 1953; Austin, 1962; for application to neuroscience, see Bennett & Hacker, 2003).

As a simple illustration of the two approaches, consider the statement, “The elliptical appearance of a coin seen from the side is a perceptual illusion”. A phenomenological response might be: “When I look at the coin from the side, I am tacitly aware of an array of possibilities – how I can move around it, grasp it, turn it around, and so on. The coin-from-the-side is not an image needing interpretation but a field of possibility.” (cf. Merleau-Ponty, 1962, p. 235). An ordinary-language analysis might be: “In ordinary talk, the words ‘appearance’ and ‘illusion’ are used in such-and-such contexts, and none of these apply in the current case. In essence, you are using these words in a narrow and eccentric manner, and your ‘statement of fact’ hides several assumptions.” (cf. Austin, 1962, p. 26; Ryle, 1949, p. 216). The two approaches have much in common. To be sensitive to ordinary language is to be sensitive to nuances of everyday life and vice versa, and it is this sensitivity that gives force to both critiques. Both the approaches have occasionally been linked to direct realism (e.g., Glotzbach & Heft, 1982; Shaw & Bransford, 1977, p. 24-25; Maratsos, 1977), but we suggest that a closer alliance is warranted. In particular, we believe that phenomenological and ordinary-language critiques are vital in developing direct realist accounts of “higher” cognitive activities such as attending, recalling, imagining, and thinking. Our analysis of language in this paper is intended to be a case study of how such critiques can be brought to bear on a specific problem.

The structure of the paper is as follows. First, we briefly summarize two direct realist approaches to speech and language, and point out their shortcomings. Next, we highlight key assumptions in traditional accounts of language and critique them using both phenomenological and ordinary-language arguments. These critiques constitute the main part of the paper. Then, we synthesize the critiques into a framework for a direct realism of communication, propose how it may be related to the perception of articulatory events, and outline some experimental approaches for its evaluation. Finally, we conclude with some reflections on the kinship between ordinary language analysis, phenomenology, and direct realism.

## Direct Realism and Language

From the start, the issue of language has been taken as a challenge for direct realism. Direct realism posits a lawful relation between the dynamics of the distal event, the nature of the medium, and the ongoing activity of the perceiver. This lawfulness tightly binds the actions of the perceiver to the structure of the immediate environment, which poses a difficulty with language. Upon the traditional view, a word is an *arbitrary* signifier of an object - there is no necessary or lawful relation between the form of the utterance “Bill Clinton” and the designated person. How to reconcile this sort of arbitrary referentiality with the preference for situated action? This tension is reflected in the debates over the nature of speech, with proposals emphasizing either the act of utterance (and therefore its situated, physical nature) or its communicative role (and therefore its referential, abstract nature).

The proposal emphasizing the act of utterance has been stated most clearly in Fowler (1986; also see Best, 1995, and Rosenblum, 2004). She suggests that speech should be treated as a structured sequence of articulatory events, with questions about the perception of speech being reformulated as those about the perception of articulatory events (Fowler, 1986). Under this view, what allows listeners to distinguish [bad] from [pad], or grasp the phonetic structure of the nonword [frIn], is that they are sensitive to the articulatory movements of the speaker such as the timing of the glottal release that distinguishes [b] from [p]. We shall refer to this as the direct perception of articulatory events (DPAE).

The proposal emphasizing the communicative role is formulated in Verbrugge (1985). He suggests that the tension between the referentiality and situated action arises from an inappropriate analogy: “The relation between word sounds and referents has been accepted as the most appropriate parallel to the relation between light patterns and objects ... An alternative ... would be to focus on the relation between word sounds and *communicative actions*, and compare this with the relation between light patterns and *events*.” (p. 163). Moreover, the notion of “event” should be broadened to include communicative intent, since “cognitive activities and affective processes are as much actions of a biological system as the overt skeletomuscular movements that the term ‘action’ usually implies” (p. 172). Consequently, language (and word use in particular) is not fundamentally ambiguous, arbitrary, representational, mediated or formal. Rather, language is to be seen as a form of social action, with words being natural indices of cognitive activities. We shall refer to this as the direct perception of communicative events (DPCE).

So what is the relation between DPAE and DPCE? On the one hand, both of them are

necessary for a complete direct realist account (after all, articulation is a part of communication). On the other hand, the relation between them is unclear. Verbrugge (1985) notes that articulatory events are not meaningful in themselves, so DPAE is “ironic, and difficult to accept, since one of the missions of ecological theory is to find a lawful basis for our perception of the *significance* of events” (p. 160). Fowler (1986) acknowledges this difficulty, but suggests that the perception of articulatory events is “one of the partitionings of an event involving linguistic communication that is perceived and used by listeners. Therefore it is an event in its own right” (p. 5). We interpret this defense as follows: When a person speaks, there are many kinds of information available for pickup – articulatory gestures, speaker identity, emotional state, hand gestures, and so on. Listeners become attuned to one or several of these kinds of information and regulate their own actions on that basis. Event theorists need to account for the specification and pickup of each of these kinds of information, and DPAE is a valid event analysis of one slice of the larger problem.

However, Fowler’s defense leaves the relation between DPAE and DPCE still in doubt. A communicative event is not several different sub-events bundled together. Rather, it is the listener’s orientation towards all of them at the same time. A formulation that treats articulatory events as a natural partition of the overall event raises the question of how the phonetic structure is linked to the talker characteristics, bodily gestures and all the rest. For example, one can tell if a friend is speaking sadly, sarcastically, or angrily, but “speak sadly” is not a characteristic of any single partition. Rather, it is characteristic of all of them together. An inferential-perception theorist would respond that these characteristics are ‘bound’ into a unified internal representation. A direct-perception theorist cannot invoke such a mediational account, and therefore the theoretical link between articulatory and communicative events remains unclear.

How then are DPAE and DPCE to be reconciled? We propose that the difficulty here is due to theoretical assumptions carried over from traditional linguistic analyses. The next five sections analyze these in greater detail, and subsequently we shall suggest that DPAE is valid in a particular *kind* of communicative context.

### Critique of Traditional Assumptions

In order to focus our critique, we shall present a “traditional view of language” which has a set of “traditional assumptions”. This is partly a strawman because there may be no single theorist who espouses all of the assumptions, but there is wide agreement on most of them. The following two quotations capture the essence of the traditional view:

Everyone who knows a language can understand what is said to him or her and can produce strings of words which convey meaning. Learning a language includes learning the 'agreed-upon' meanings of certain strings and learning how to combine these meaningful units into larger units which also convey meaning. (Fromkin & Rodman, 1978, p. 163).

The capacity to use spoken language boils down to an ability to assign meaning to sequences of speech sounds. It thus involves a harnessing of two different types of knowledge. One interacts with the central conceptual system and controls how we compute meaning. The other is phonological and regulates the issuing of instructions to the motor-perceptual systems for the purposes of producing and recognizing speech sounds. (Harris, 1994, p. 1).

The assumptions underlying the above views may be summarized as follows. (1) Each person has a private internal realm (his or her mind) that is populated by thoughts. Upon some occasion, the person decides to reveal a thought, and the linguistic utterance is a public sign of the private thought. (2) Listeners recover the thought from its public sign. Being fundamentally sealed off from the speaker's mind, listeners cannot truly know the thought behind the utterance. (3) The speech utterance is composed of meaningless phonological units. (4) A sequence of phonological units (a word) gains meaning through conventional association. (5) There is a principled difference between the knowledge of linguistic structure and the exercise of linguistic abilities.

There is some debate within the cognitive tradition over some of the above assumptions. For example, connectionist, cognitive-functional and speech-act theorists (McClelland & Elman, 1986; Gibbs, 1994; Clark, 1996) challenge assumption 5, and exemplar theorists such as Bybee (2001) challenge assumption 3. We suggest that a more radical surgery is needed – all five of the assumptions are flawed and should be discarded. They are mainly the result of a too-narrow focus on certain linguistic phenomena and on certain ways of describing the phenomena. Since assumption 5 has already received due attention, we will focus below on assumptions 1-4.

### Speaking is Not Separate From Thinking

In the traditional view, the relation between thought and speech is generally conceived in one of two ways. One is that the thought precedes the utterance, with a “decision to speak” determining whether the thought is “let through” into speech. The other view is that thought does not precede the utterance but rather runs parallel to it. Here the decision to speak is more akin to opening a channel between two streams. Common to both these views is the notion that thinking and speaking are distinct cognitive activities, with the “decision to speak” acting as a gatekeeper.

We shall argue below that this notion is incorrect.

Before we proceed with our critique, we should clarify what we mean by “thought”. There is an active debate in cognitive science whether thought is based on propositions, imagery, or language (or some combination thereof). Addressing this issue would go beyond the scope of this paper. For current purposes, we assume that there are both nonlinguistic and linguistic modes of thinking. In this paper, we restrict ourselves to claims about *linguistic* thoughts, i.e. the kinds of thoughts that are expressed in speech.

*Speech is Not Preceded By a “Decision to Speak”*

Is speech always preceded by a deliberate decision? This is certainly not the case in everyday spoken conversation. When I encounter a friend, I no more “choose to speak” than I “choose to smile”. Greeting that friend and chatting are not preceded by deliberate decisions, and in fact one of the joys of meeting a friend is a relief from having to make such decisions. Even when decisions about speaking are made, they often have no force. I decide not to tell my friend some scandalous gossip but once I see him, I blurt it out anyway; I decide to tell him some unwelcome news, but then find myself unable to say it; I snap at him and immediately regret it. Situations such as these indicate that in everyday conversation the decision to speak consists of the act of speaking itself.

One alternative here is that the “decision to speak” is not deliberate, but rather made by a subconscious cognitive process. But this is untenable. A stone cannot decide to fall down and a heart cannot decide to pump blood; in order to say that one can make a decision, it must also be possible to say that one can be decisive, determined, resolute, or indecisive. These cannot be ascribed to the stone or the heart, and neither can they be ascribed to the cognitive process. More generally, one cannot take terms ascribed to a *person* in a social context – making decisions, being decisive or indecisive – and ascribe them to a particular part of that person (Bennett & Hacker, 2003, call this “the mereological fallacy”).

The import of the above points is that speech is not preceded by a decision to speak. Rather, the speech depends on the situation and the audience – in some contexts, we find ourselves talking freely while in others we find ourselves tongue-tied. Two objections may occur to the reader at this point. First, one can mentally rehearse a sentence and decide to utter it upon some occasion. Second, in many situations one can prevent oneself from talking (by suppressing a retort, for example). What is going on in these cases? We suggest that they may be understood

by comparing speaking to smiling. Most of the time, peoples' smiles are spontaneous and elicited by the situation. Sometimes one forces a smile, but this is more than "turning on" the smile; one has to cultivate the muscle control for mimicking a sincere smile. Sometimes one suppresses a smile, but this is more than "turning off" the smile; one has to actively hold back the smile by biting one's lip or grimacing. In general, forcing and suppressing are sophisticated skills that are superimposed on basic spontaneous smiling, and the same is true of speech. It is of course important to understand how deliberate control over one's speech is possible, but it is also important to note that such speech is rare in everyday conversation.

### *Speech is Not the Product of a Thought*

The traditional view holds that thinking and speaking are distinct cognitive activities, and that behind every speech utterance stands a thought that is responsible for the syntax and word choice of that utterance. This position has several problems. First, it is inconsistent with everyday usage. To speak thoughtfully is to think carefully about an issue before speaking, as when one is asked to give an opinion on an important matter. But to speak thoughtlessly is not to utter meaningless sounds, it is to speak in a foolish or impulsive manner. That is to say, thoughtfulness is an attribute of meaningful speech and not a cause of it. Second, positing a thought behind each utterance leads to conceptual difficulties.

*Scenario 1.* A person, cut off in traffic, leans out of the car and shouts angrily, "Are you blind?! Moron!!". Before shouting, did he have a thought about whether the other driver was blind? Was the anger *part* of his thought? If yes, did he learn to translate the anger-thought into the angry tone and expression of the actual utterance? If no, then does the *same* thought stand behind angry, sarcastic, and solicitous utterances of "Are you blind"?

*Scenario 2.* A person is cut off in traffic as before, but now he has a passenger he wants to impress. Instead of shouting out the words as in Scenario 1, he silently says them to himself (i.e. he curses to himself). Should we say that there are here a set of precursor thoughts (the same as in Scenario 1) that give rise to another set of thoughts (the silent curses)? If there is only one set of thoughts – if the precursor thoughts and the silent curses are one and the same – should we say that a person always silently curses to himself *before* cursing out loud?

*Scenario 3.* A person is recounting a funny incident to friends, and finds it so funny that her speech is interspersed with her own laughing and chuckling. Are her speech-thoughts interrupted by her laughter-thoughts? Or her speech-thoughts by her laughter? Or is the laughter *part* of her

speech-thoughts?

We submit that the questions posed above have no coherent answers, and that the problem is not a lack of knowledge but rather the underlying assumption, viz. that every utterance is backed by a thought. The above scenarios and related situations – greeting, cursing, flirting, joking, exclaiming, snapping – are the vast majority of our speech interactions, and it is noteworthy that the traditional view is unable to give a coherent account of them. It is also unable to explain why, in situations like Scenario 1, the speaker feels the need to express his thoughts at all. As Merleau-Ponty notes, “If speech presupposed thought ... we could not understand why thought tends towards expression as towards its completion” (Merleau-Ponty, 1962, p. 206).

One possible response to our criticism is the following: “The ‘thought’ is not a conscious formulation but a subconscious one – an internal representation of the meaning of the utterance that causes the actual utterance”. But this really will not do. Consider Scenario 1: If the angry utterance is caused by the internal representation, then is the anger part of the internal representation? If yes, it has to get translated into the expression of anger in speech (as in Scenario 1) or into another internal representation that signifies “anger” and accompanies the cursing-to-oneself (as in Scenario 2). Moreover, there are many different manners of anger-expression speech – one can scream, hiss, speak coldly, formally or bitingly, speak in an uninflected monotone, and so on – so either the internal representation has to mark all the nuances of the expression or else the translation mechanism has to somehow select a context-appropriate expression. At this point the idea that an internal representation stands behind every emotional utterance is reduced to homuncular absurdity.

### *Speech is a Form of Thought*

The import of the above two sections is that speech can be thoughtful, but it is not a product of thought, nor is there a “decision to speak” that separates thought from speech. In short, speech is a kind of thought (cf. Merleau-Ponty, 1962, p. 209).

The notion that thinking can be constituted by activity is easier to grasp in other domains. A jazz musician may be said to play thoughtfully if, for example, his music surprises the audience in clever ways or explores interesting chord relationships. To a perceptive listener, the thoughtfulness does not have to be inferred from the play; rather, it is evident in the music (Ryle, 1949, p. 32). The music *is* the musician's thought at that moment, and it is the unique intersection of the musician's skill (his technical expertise and repertoire) with that particular context (the

piece that he is playing, the audience, the other players in the band). The same is true of any other activity involving skill and expertise. A thought is skill-in-context, and each activity makes possible its own characteristic kind of thought. In particular, a speech utterance is speech-skill-in-context. It constitutes the speaker's thought at that moment, and its “meaning” is its functional role in that context (we shall address “meaning” in more detail later on).

Three points may be noted regarding the above view. First, it fits naturally with our earlier comments regarding the “decision to speak”. We might decide to think about some topic or other, but we do not decide to have a particular thought nor do we plan a sequence of thoughts. Thoughts occur to us unbidden and so does our speech. Chafe (1982) notes that “[spontaneous spoken language] is produced in spurts, sometimes called idea units ... [They] typically have a coherent intonation contour, they are typically bounded by pauses, and they usually exhibit one of a small set of syntactic structures” (p. 37). Second, if speech is a form of thought, then the “content” of an utterance cannot be separated from the particulars of the utterance itself. This point can be appreciated by anyone who has ever heard a good joke ruined by a bad delivery, and it is true more generally. *All* the aspects of the utterance – whether it is shouted or whispered, spoken quickly or slowly, crisply enunciated or slurred, as well as the body posture, hand gestures, and facial expressions of the speaker – are part of the thought being expressed. Third, a person is capable of non-linguistic thinking, and also of linguistic thinking without speaking. However, the difference between thinking-*without*-speaking and thinking-*through*-speaking is not simply the act of speaking, as if thinking in both cases was the same. One can practice dribbling a basketball by oneself, but during a game one is not doing two things (dribbling by oneself and also playing the game). Rather, in the game dribbling a basketball is something one does in relation to other players. Likewise, thinking-*without*- and thinking-*through*-speaking are different ways of linguistic thinking – individually in the former case and communally in the latter.

Our point can be put another way. It is a truism that speech is a form of communication. However, the historical meaning of “communicate” was not “transmittance of a message” but rather “to make common” (cf. “commune”). Speech *is* communication, but what is made common is not sounds or articulatory movements but thoughts. Speech is the public sharing of thoughts.

### *The Speaker's Thoughts are Shaped by the Listener and the Situation*

A central tenet of the traditional view is that the speaker initially has a thought *X* that can (in

principle) be about anything at all. It is because of this sheer variety that speech is supposed to require compositionality of structure and reference, so as to enable the listener to reconstruct the same *X*. However, thoughts rarely occur *de novo* to the speaker. More often, they are shaped by the speaker and the listener's shared historical and immediate context.

*The historical context.* Consider a simple situation. A asks B, "Can you pass the salt?". It is not obvious what this means. Did A mean it as a question, a request for salt, an insult ("Can you pass the salt?"), or a sarcastic comment? Did A ask haughtily, politely, or warmly? Did he get B's attention by calling B's name or by snapping his fingers? All of these linguistic and situational factors make possible an immense variety of meanings. Assume, for example, that A meant it as an insult. But how did A know to insult in just *this* particular way, and that his insult would be understood by B? The answer is the shared history of A and B, their familiarity with each other's thoughts and manner of speech, and their familiarity with dinner-table conventions. As a result of this history and familiarity, A does not have to think about the insult beforehand – to him, at that instant, that particular insult (with that specific tone of voice, gesture, etc.) presents itself as the most natural thing to do. That utterance *is* his thought at that moment, and that utterance/thought is the intersection of that shared history with that situation.

*The immediate context.* The immediate activity of the listener also shapes the development of the speaker's thought. In face-to-face conversation, for example, the speaker is sensitive to a particular pattern of eye movement from the listener, and he is immediately aware if this is violated, e.g., if the listener unexpectedly avoids eye contact or stares fixedly at the speaker's eyes. Such unexpected eye movement can cause the speaker to lose his "train of thought". For the exchange to play out normally, the speaker and listener have to perform a finely-timed ballet of looking at and away from each other. A similar sort of ballet occurs in terms of head and body movements, vocalizations like "umm" and "uh-huh", and (we surmise) at key intra-sentential points like focus phrases and ends of intonation contours. Thus, the listener-activity cannot be arbitrary or random; it is *of necessity* coupled with (and complementary to) the speaker's fine-grain perceptuomotor activity and other relevant activity in the vicinity. "The phonetic 'gesture' brings about, both for the speaking subject and for his listeners, a certain structural co-ordination of experience, a certain modulation of existence, exactly as a pattern of my bodily behavior endows the objects around me with a certain significance both for me and for others." (Merleau-Ponty, 1962, p. 225).

The upshot of the above two points is that the speaker's thought is not constructed *de novo*; rather, it is shaped by shared history of the speaker with that listener in that situation, and by the listener's immediate activity. When we encounter a close friend, that friend's very presence

opens up to us an array of possible meanings – we suddenly *think*, listen, smile, groan, and laugh in certain ways that are unique to the history of interactions and situations that we share with that friend. Likewise, the presence of a disliked person opens up an array of possible meanings unique to our history with *that* person.

### Listening is Not A Prelude to Understanding

In the traditional view, the activity of the listener is held to be symmetric to that of the speaker: The utterance prompts the listener to construct a thought in her own mind. If this reconstructed thought is similar to that in the speaker's mind, then the listener has understood the utterance and can formulate her own thoughts or actions in response. However, the listener can never know *for certain* whether the reconstructed thought is really the same as that in the speaker's mind. As we shall argue below, both these positions (reconstruction of the thought, and uncertainty about it) are unjustified.

#### *The Listener Does Not Reconstruct the Speaker's Thought*

It is a commonplace that speech can lead to thoughts in the listeners. After listening to a compelling argument, we rethink our position; after hearing an evocative description, we visualize the described scene. However, these are the listeners' *responses* to the utterances. The traditional view asserts that before any such response-thoughts can occur, the listener has to first recover (or reconstruct) the speaker's thought. We suggest that (1) the listener's understanding of the utterance is shaped by his perceptuomotor activity, and (2) this activity precludes any sort of thought-construction. Our argument consists of four observations.

*There is no cognitive distinction between literal and figurative understanding.* It is often held that listeners first have to understand an utterance literally, and then use this to understand it figuratively (e.g., Jackendoff, 2002). However, experimental evidence suggests that the literal understanding is not primary. The data are extensively reviewed in Gibbs (1994), but a brief example will give the flavor. In one experiment, subjects first read an idiom, e.g., *He kept it under his hat*, in a context that biased either a literal or an idiomatic understanding. Then, the subjects read a sentence that paraphrased either the literal meaning (*It's beneath his cap*) or the idiomatic meaning (*He didn't tell anyone*), and had to respond whether that sentence was grammatical or not. The results indicated that when the subjects read an idiom in an idiomatic context, they were as fast to respond to the idiomatic paraphrase as to the literal paraphrase

(Gibbs, 1994, p. 95). Based on results such as these, Gibbs concluded that “as long as the contexts are equally explicit, the same utterance can be used in any one of a variety of pragmatic roles (e.g., literally, metaphorically, sarcastically, or as in indirect request) without significantly affecting the manifest difficulty of processing” (Gibbs, 1994, p. 110). We draw a similar conclusion, viz. that there is no cognitive distinction between literal and figurative understanding. The listener understands a literal request (“Can you help me?”) and a sarcastic indirect request (“Thanks for your help!”) in the same way, with neither of them being more fundamental or more basic than the others.

*The listener’s understanding of an utterance is identical to his understanding of the situation.* In order to understand a sarcastic comment or an indirect request, the listener needs to understand the overall situation (comprising the setting, the task at hand, the speaker’s goal, and so on). The same situation-understanding is also needed for literal utterances, for two reasons: (1) As noted above there cannot be a cognitive distinction between figurative and literal understanding. (2) In many situations, the speaker can make either a figurative or a literal comment, and it is only by being alert to the actual situation that the listener can distinguish between the two. We suggest that in general the listener’s utterance-understanding is identical to his situation-understanding. It is akin to a patch of color in a painting. There is only one kind of meaning, and that is what the painting is about (is it about water-lilies? sunsets? faces?). The color patch contributes to the meaning of the painting, but it does so by the manner of the brush stroke and the shade and texture of the color, not by contributing a tiny bit of painting-meaning. To understand the patch (the utterance) is to understand the painting (the situation), and vice versa.

*The listener’s understanding of a situation is constituted by his activity.* How does one “understand a situation”? An example may help clarify this notion. I am walking in a dark alley and see a person slowly walk into my path. I become alert and cautious, walking quickly, tensing myself to run if necessary, looking carefully for potentially threatening movements and for possible escape routes, listening for sounds which might betray an accomplice, and so on. These activities involve perceptuo-motor engagement at very fine time scales, such as the sequence of visual saccades and sensitivity to binaural disparities. This ensemble of gross- and fine-scale activity – how it is patterned, how it evolves, and what it prefigures – is my way of relating to that particular situation, and it *constitutes* my understanding that the situation is threatening. Indeed, it makes no sense to divorce the activity from the understanding. One cannot find a situation threatening and at the same time be tranquil or act in a way inappropriate to that particular quality of threat. Put simply, there are not three events (I perceive the objective situation, interpret it, then respond), but rather there is only one event (I am engaged in a

particular mode of activity).

*The listener's understanding of an utterance is constituted by his activity.* As I walk forward in the dark alley, the person extends his arm and says [dʒhævðəraɪm]. How to understand this? This depends on his stance, gaze direction, the precise nature of his arm extension (how it is extending, how the hand and fingers are flexing, etc.), tone of speech and – most importantly – on how I attend to all these aspects. These factors shape my own posture, manner and direction of gaze (are my eyes focused on one point or saccading between several foci?), and manner of listening (am I attending to other sounds in the vicinity?). Thus, his gesture and utterance shape my mode-of-activity. This elaborated mode-of-activity constitutes my understanding of the situation, and I understand the gesture and utterance *as part* of that larger understanding. For example, if I perceive his arm-extension as the act of bringing up the wrist, I may understand the utterance to be, “Do you have the time?”. Conversely, if I perceive it as the act of extending his palm, I may understand the utterance to be, “Do you have a dime?”.

In summary, an utterance is inseparable from its situation; to understand the situation is to be engaged with it; therefore to understand the utterance is to be engaged with the situation. If the understanding involved the construction of a thought, such a thought has to organize all the fine-grain perceptuomotor activity of the listener. At the same time, it has to be a reconstruction of the *speaker's* thought. In other words, the reconstructed thought has to replicate the speaker's point of view while being wholly integrated with the listener's activity. It is this conflict that makes thought-reconstruction untenable. The listener has a particular historical and situational perspective, and he can no more “reconstruct” the speaker's perspective than he can be in two places at once.

### *The Listener is Not Fundamentally Uncertain About the Speaker's Thought*

A second component of the traditional view is that understanding speech is an error-prone activity, akin to guessing the contents of a black box from superficial signs. However, once speech is seen as a form of thought, there is no question whether the listener is shut off from the speaker's true intent. The thought of the speaker is no more hidden from the listener than the thoughtfulness of a musical performance is hidden from the audience (cf. the debate over “direct” versus “indirect” perception, Austin, 1962, p. 15; Shaw & Bransford, 1977, p. 24-25). To grasp the thoughtfulness of the performance is to attune oneself to the context of that performance (be familiar with the style of playing, anticipate how the performance is going to unfold, and so on). Likewise to apprehend a thought is to attune oneself to the context of that

thought – it is to follow along bodily, verbally, and mindfully in a manner appropriate to that thought.

The above point can be made in another way. Consider a young child being scolded by her parents. For the child, the intent of the parents is completely defined by the situation and utterance – what it means for a parent to be angry *is* for him or her to say (in the appropriate context), “Go to your room!” or “Go take a time out!”. The child no more wonders whether the parent is *really* angry than she wonders whether the entity called “cat” is *really* a cat. Intents and thoughts are as plain in the child’s world as cats and tables, so the question of uncertainty does not even enter the picture. Of course, with experience, the child learns to differentiate between different kinds of anger (genuine vs. suppressed vs. feigned) just as she learns to differentiate between different kinds of cat (kitten vs. adult cat vs. toy cat). However, the differentiated intents are still plain in her world; the parents’ behavior and context are subtly and systematically different for the different kinds of anger, and the child is sensitive these differences. In short, once the child has attuned herself to the context of the utterance/thought, there is no question of uncertainty about the intent.

One might here object, “If the speaker’s intent is always plain, how can lying and miscommunication be possible?” The response to this is similar to the explanation of mirages in direct perception: “The realist position is ‘as things are perceived *so they are*’ – which is to be distinguished from the claim ‘as things are perceived *so they really are*’.” (Heft, 2001, p. 81; also Fowler, 1991, p. 2914). If you see a marching band on the street with floats and balloons, you are not seeing the consequences of a hidden parade-event, you are seeing the parade itself (cf. Shaw & Bransford, 1977, p. 36-37). If you are told, “The organizers did not want to have a parade, they simply told this band and these floats to move in this particular way. Therefore, this is not really a parade!”, you would be right to respond, “That’s a real band, and those are real floats and balloons. This *is* a parade, possibly one put on for strange motives, but a parade nonetheless”. Likewise, if you see a man shouting furiously, there is no “thought of fury” (hidden parade-event) that causes the pattern of action (the elements of the parade). The man may be pretending (putting on the parade for strange motives), but he is not mimicking the consequences of a fury-thought, he is pretending to *be* furious (it *is* a parade). In the same way, the liar is not shamming the bodily consequences of a thought, he is shamming the thought itself. When lying successfully, the liar is not holding on to the true-thought while moving his articulators to produce the lie; rather, in uttering the lie he is briefly taking on the false-thought (indeed, the most successful liars convince themselves for the nonce that they are telling the truth).

Now, suppose that part of the parade consists of a long line of cars. You overlook the other elements of the parade, only see the line of cars, and mistake the parade for a motorcade. However, your mistake is *not* that you thought the spectacle was due to a hidden motorcade-event rather than a parade-event; rather, your mistake is in overlooking those parts of the spectacle that distinguish a motorcade from a parade. Likewise, to misunderstand a speaker (e.g., take a sarcastic remark at face value, or mishear a mumbled comment) is not to infer an incorrect thought, but rather to mis-attend to the thought.

*Listening is an Active Relation of the Listener to the Speaker*

A key implication of the above arguments is that speech understanding requires the multimodal perceptuomotor engagement of the listener. An objection might be raised here that listening is possible in two seemingly non-interactive situations: (1) face-to-face conversation with eyes closed and body motionless and (2) listening to pre-recorded speech over headphones. A full treatment of these objections will take us beyond the scope of this paper, but below we sketch the outlines of our response.

In the first case (face-to-face conversation), the main point to note is that closing the eyes and keeping still are part of the interaction (for example, they may indicate the listener's irritation or interest). Closing one's eyes during a conversation is akin to covering them with one's hand; the action *masks* the activity of the eyes but does not render them inactive. Likewise, keeping still during a conversation is very different from being motionless while resting. Consider a listener who is being insulted but "holds himself back". Here, the listener's tendency is to retort or somehow stop the speaker, but he *actively counteracts* his own tendencies (e.g., "bite one's tongue"). That is, the listener's immobility actually involves covert muscular effort that masks his engagement with the speaker and situation.

For the second case (listening to pre-recorded speech) consider first the following situation: I see a video of a person silently gesticulating. His actions seem haphazard and meaningless, but suddenly I realize that he is miming how to iron a shirt. His actions fall into place because I *see* them as part of a larger situation (Wittgenstein, 1953, §539). But how is this possible without an actual shirt or iron? The answer is that in seeing a situation "as" something, one relates to (or interacts with) the situation in a certain way. With experience, the relation can be evoked with only a fragment of the original situation, such as the actions of the mime. Something similar occurs in listening. To listen to someone speak is to relate to that person and that situation. With experience, this relation can be evoked with only the speaker's voice. Put simply, the listener

“detaches” himself from the current situation and “places” himself in the situation of the speech. This has three implications. First, when listening to prerecorded speech, the listener always has a sense of the speaking situation (e.g., the age and sex of the speaker, his or her movements while speaking, and the location for the speech). Second, the “detaching” is a learned skill specific to particular situations. For example, one can follow an audio-book in a quiet bedroom but have difficulty following it when on the bus. Finally (and most importantly), the detachment is never complete. If one listens to a story over headphones while strolling in a park, the park becomes intertwined with the understanding of the story (and in recalling the story later on, it will be difficult to separate the two). The listener’s interaction involves a fusion of his immediate situation and the situation of the recorded speech.

### The Utterance is Not Composed of Phonological Units

Much of the force of the traditional view of language comes from the enormous amount of detailed data on phonology, lexical morphology and grammar (e.g., Bloomfield, 1933; Harris, 1994; Jackendoff, 2002). Much of this data is based on the view that languages have a “duality of patterning”, “the property of human speech by which a small number of meaningless elements (the phonemes) can be combined into a limitless number of meaningful expressions (words and sentences)” (Trask, 1996).

At one level, “duality of patterning” is the innocuous observation that there is no fixed relation between sound and meaning. However, it has also been put forward as a claim about cognition, namely, that speech production involves the translation of a word into its constituent “meaningless elements”, and that speech perception involves the reverse process. We suggest that this claim involves problematic assumptions about what constitutes a word, and about the relation between linguistic and indexical properties.

### *A Word is Defined by Function, Not By Acoustic Structure*

There is a peculiar assumption that underlies the claim of duality, and by extension, the notion of a word. Consider how Bloomfield (1933) justifies the notion of a phoneme:

The phonetician finds that no two utterances are exactly alike. Evidently the working of language is due to a resemblance between successive utterances. Utterances which in ordinary life we describe as consisting of “the same” speech-forms – say, successive utterances of the sentence *I’m hungry* – evidently contain some constant features of

sound-wave, common to all utterances of this “same” speech-form. Only on this assumption can we account for our ordinary use of language. (Bloomfield, 1933, p. 76, §5.2).

Note the term ‘resemblance’. A layperson, upon hearing two people say, “I’m hungry”, might comment, “They are both hungry”, or (if interrogated some more) “They both said the same thing”, or (if interrogated even more), “They both spoke the same way”. However, no amount of questioning would elicit the response, “There is a resemblance between what they said”. In fact, to even attempt to do so – “Ignore the meaning and the differences in voice, loudness, pitch and timbre. *Now* is there any resemblance?” – verges on the nonsensical. One does not ordinarily talk about the acoustic shapes of utterances, but only of their role. Bloomfield assumes that the meaning-resemblance of the utterances *must* be supported by a sensory-resemblance. The conceptual flaw here may be seen by an analogy to tool use.

Consider the activity of digging. A person observes that digging can involve many different tools – shovel, pointed stick, sharp metal bar, a spoon, one’s hands, etc. – and decides that (a) there is some hidden physical property *P* common to all these tools, (b) it is this property *P* that is associated with digging, and (c) it is possible to identify *P* by detailed physical comparison of the tools. This project is of course doomed to failure, because there is no single act of digging. There is rather a *family* of digging-activities – one can dig a grave or a tunnel, dig through rock or sand or mud, dig through garbage or clothes in a drawer, etc. – and different tools are appropriate in different situations. What makes them all *digging*-tools are similarities in the manner and context of their use, not a common physical property. When one takes a group of tools that are common in *function* and looks for a commonality in *form*, the presumed commonality *P* will have a peculiar status – it will appear to be an immanent physical quality that is not identical to any single tool, but yet is not identical to the function. Simply put, *P* is a reification of the function and it is a mistake to set about looking for it, or worse, to speculate how it causes the function.

The upshot of the above discussion is the following. Many different utterances, spoken by different people in different contexts, can all be transcribed as the same word. However, the utterances do not have to share a common acoustic feature which is in turn linked to the common meaning. There are only concrete utterances in concrete situations. It is notable here that speakers in a non-literate society find our notion of a word to be incomprehensible (Lord, 1982).

*Linguistic and Indexical Information Cannot be Separated*

When hearing someone talk, listeners are sensitive to “what is being said”, as well as to the identity, gender, size, age, health, social status of the speaker, and so on. The traditional account of this ability goes something like this: The incoming auditory signal is simultaneously accessed by many different processes. One process identifies phonological units, another the speaker identity, yet another the gender, and so on. Each process factors out the variability introduced by all the other kinds of information and also performs its basic recognition function. The outputs of all these linguistic processes are sent to a separate semantic process that evokes the actual meaning of the utterance. This scheme is conceptually problematic, for four reasons.

*The different streams of information cannot be considered separately from each other.* For example, familiarity with a speaker’s voice can help with linguistic identification (Nygaard & Pisoni, 1998); social status and age analysis can help identify the speaker and the linguistic content; the linguistic content can help identify the social status (e.g., by the use of prestige terms). Any sort of modular scheme, on the other hand, creates the very “problem of variability” that it is designed to solve.

*The phonological details cannot be considered separately from the context.* Central to the traditional scheme is the notion of some kind of reduction – some of the variability of the auditory information is thrown away to produce a robust, orderly sequence of phonological units. But what information should be thrown away? The ‘suprasegmental’ aspects of an utterance such as the intonation, loudness, and rate are all informative. Moreover, the phonetic detail of the ‘segment’ itself varies systematically with linguistic context (Local, 2002; Bybee, 2001) and sociolinguistic context (Chambers, 1995). Context-dependent variation cannot simply be factored out by the perceptual process, because the variation is part of the information to be perceived.

*Listeners have a unified perceptual experience of the speech utterance.* When we hear an utterance, we do not have separate experiences of the different information streams. Rather, we always hear an utterance as an utterance made *by someone*, and we are at the same time aware whether the person is male or female, whether they sound attractive, whether they have an accent, what their social class is, and so on. At a party, for example, we do not confuse the gender or identity of one speaker with the accent of another. Such a unified percept poses a dual problem for the modular scheme. First, all the different streams have to be integrated at some point to produce the unified semantic representation. Second, the integration has to preserve all the ‘flaws’ in the auditory input. For example, if the utterance is spoken with a lisp or a nonnative

accent, we are aware of the lisp or accent (i.e. the infelicities are not 'cleaned up').

*A separate semantic process is untenable.* If the semantic processing is to be separate from the speech processing, the interface between them has to be fairly restricted. That is, the speech has to be transformed into some canonical form before being passed along for semantic processing. However, the notion of a canonical form is implausible. First, the semantic importance of fine phonetic detail – and the preservation of infelicities – implies there cannot be any reduction or 'filling in' (also see Pessoa, Thompson, & Noë, 1998). Second, the understanding of the utterance is contingent on the perceptuomotor activity of the listener. Consequently, it is difficult to make a clear distinction between speech and semantic processing.

Some of the above problems have been noted before. For example, McClelland and Elman (1986) proposed an interactive activation model to explain how variability can help rather than hinder processing. Likewise, Johnson (1997) and others (Pierrehumbert, 2002; Hawkins, 2005) have proposed exemplar models to accommodate the effect of fine phonetic detail. An analysis of these approaches is beyond the scope of this paper. For current purposes we simply note that the approaches, while innovative, still draw a sharp distinction between speech and semantic processing, and are therefore susceptible to problems 3 and 4.

Our own proposal is simply this. The idea of phonological units stems partly from the notion that the speech utterance is initially meaningless and has to be *given* meaning by the listener. But as we have argued earlier, speech is a form of thought and the listener is directly sensitive to the meaning of the utterance/situation. There is no point at which the speech utterance is non-meaningful. An utterance is not like a letter sent from A to B; it is more like a warm hug or a sharp slap. In the proper context, there is no question of 'interpreting' the hug – it is not an intrinsically-meaningless sign of affection, it *is* affection. Likewise, in the proper context, there is no question of 'interpreting' the utterance, and understanding it does not involve an initial parsing into phonological units such as phonemes, syllables, or articulatory gestures. Moreover, since the utterance-understanding is inseparable from the situation-understanding, the "linguistic" information cannot be separated from the "indexical" information.

### The Meaning of an Utterance is Shaped By its Use

We have so far used the term 'meaning' rather loosely. We have asserted that the speaker's utterance is a form of thought; this thought is part of the situation; the listener follows the thought/situation; and the meaning of the utterance/thought is its function in that situation.

However, it may still appear puzzling how an utterance gets its arbitrary meaning: How can one sit comfortably in an office, talk about tigers and galaxies, and be understood? How can the meaning of *galaxy* be constituted by any sort of action? We shall try to address this puzzlement by placing our work within the larger philosophical debate.

In the philosophy of language, many theories assume that the utterance is simply associated with its meaning, and that the real issue is the kind of entity that ‘meaning’ is. For example, is it an object in the world, the operant elicited from the listener, some kind of genetically-endowed innate knowledge, or a cultivated internal representation (see Jackendoff, 2002, chap. 9, for an overview)? Our position is close to that formulated by Wittgenstein (1953): “For a large class of cases ... the meaning of a word is its use in the language” (§43). This formulation has proved controversial, and it is instructive to note one general criticism of it:

This view [of meaning] is that there is no fixed meaning associated with linguistic expressions; rather the best one can do is catalog the contextual uses of expressions.... The message conveyed by an expression is indeed heavily influenced by one's understanding of the context. But on the other hand, the expression must convey *something* with which the context can interact. If it did not, a hearer could in principle know from the context what message was intended, without the speaker saying anything at all! It is important to factor out the respective contributions to understanding made by linguistic expressions and by context; this cannot be done by focusing on context alone. (Jackendoff, 2002, p. 280; emphasis in original).

Jackendoff’s main concern is that an expression has to convey a mentalistic entity of some sort (thought, idea, image, representation); if it is entirely dependent on context, then it is superfluous. But is this true? When friends meet, they typically say “Hello!” to each other. The greeting is predictable from the context, but one cannot invoke this predictability and refrain from the greeting (that would convey another meaning entirely). Or, two strangers in an elevator chat about the weather. The back-and-forth of the chat is mundane and routine, but again, one cannot invoke this fact and say silent. Or, when a group of old friends meet, the conversation topics, jokes, and reactions are typically quite predictable, but one does not (cannot) stay silent. The same principle extends to the structure of the utterance. One may ask a friend, “Where’ll we go for dinner?”; in this utterance there are predictable grammatical elements such as ‘will’ and ‘for’, but the speaker cannot leave them out; the friend may reply [tʃar'ni:s], in which case the [ni:s] may be predictable, but the speaker cannot leave that out either.

Based on these examples, we can turn Jackendoff’s concern on its head: why, in these and other routine situations, do speakers speak and listeners listen? The answer is that the talking and listening is part of the situation and *completes* it. We cannot help tapping our feet to rhythmic

music, and likewise we cannot help situation-specific talking and situation-specific listening. These are not merely actions imposed from without or routine imposed from within, because we actively seek out situations where we can “fall into” such activity (such as chatting with friends). Moreover, the situation is not a static background to the conversation; rather, the ‘linguistic’ and ‘paralinguistic’ content of the utterance changes and elaborates the situation. Just as the meaning of a patch of color is how it fits into the overall painting, the meaning of a gesture or a wordsound is how it fits into the situation. More generally, the meaning of an *isolated* color patch in a painting (e.g., a splash of scarlet may have a connotation of danger or of bloodiness) is the ensemble of paintings in which that color patch has occurred, and likewise the meaning of an isolated wordsound is the ensemble of situations in which that wordsound has occurred. Put simply, *meaning is constituted by the context-appropriate activity of the participants.*

This, then, is our reading of Wittgenstein’s dictum that “the meaning of a word is its use in the language”. Wittgenstein is not evicting meaning from the mind ; rather, he is placing both mind and meaning in the world. Meaning is use because use (the linguistic activity) is always mindful (context-appropriate and skillful). Moreover, because meaning is constituted by public activity, one can no more have a “private” meaning for a word than one can play a “private” game of tennis (cf. Wittgenstein’s argument that one cannot have a “private language”, explicated in Bennett & Hacker, 2003, p. 97-103; also McDowell, 1998, Part 3).

One consequence of the “meaning as use” view is that utterances do not have sharply bounded, context-free meanings. For example, a person’s eyes may be described as round, blue, or good. The first use refers to the shape outlined by the eyelids, the second to the irises, and the third to eyesight; there is no single core meaning of “eyes” that makes possible all these uses (Suzuki, 2001, p. 45). Likewise, there are many ways to use the verb “hold” – hold a hand, hold a pen, hold a mattress, hold someone, hold someone back, hold someone’s gaze – and here too there is no single core meaning. How then does one learn about “eyes” and “hold”? We suggest, following Wittgenstein, that one learns the *situations* in which these terms occur, where a ‘situation’ includes both the verbal context (e.g., the manner and structure of the utterance) and the nonverbal context (e.g., the presence of the person being described). To know the meaning of “eyes” and “hold” is to master their respective situations of use. Once a language is mastered, the speaker does not “search” for words because the appropriate words simply present themselves. As Merleau-Ponty (1962) notes, “I reach back for the word as my hand reaches towards the part of my body which is being pricked” (p. 210).

A second consequence of “meaning as use” is that the relation between a word and its meaning is not arbitrary. The traditional view distinguishes between natural signs (such as

scowling when angry) and arbitrary signs (the relation between term *cat* and the animal); the latter is held to be arbitrary because there is no necessary relation between the word and the referent. But this is phenomenologically incorrect. While it is conceivable that English speakers could call a cat a *neko*, it is a plain fact that they do not. For a child learning English, the words are used in a contextually consistent manner, so the wordsound *cat* is as much a part of a cat as four legs and a tail (also see Verbrugge, 1985, p. 180). For an English speaker, the word *disgust* is itself vaguely disgusting, and the thought of using it to describe something succulent feels awkward. To again quote Merleau-Ponty, “It is no more natural, and no less conventional, to shout in anger or to kiss in love than to call a table ‘a table’ ... It is impossible to superimpose on man a lower layer of behavior one chooses to call ‘natural’, followed by a manufactured cultural or spiritual world” (Merleau-Ponty, 1962, p. 220).

A final consequence of “meaning as use” concerns linguistic structure such as grammar and morphology. It is typically held that such structure is either due to intrinsic rules (e.g., Pinker and Ullman, 2002) or to extrinsic probabilities of occurrence (e.g., McClelland & Patterson, 2002). We suggest a third option: linguistic activity is structured because the *situations* that we encounter are structured. When one meets someone new at a party, in some ways it is like encountering a stranger on the road, in other ways it is like meeting an acquaintance; in some ways the person is familiar (his face is similar to that of another friend), in other ways he is unfamiliar (but his hair and eyes are different!); and so on. An array of possibilities opens up – what to say, how to move, what to do – that are not arbitrary but arise from one’s prior activity in situations like this. Linguistic regularity is likewise structured. When one hears the nonword *frink* in a context of blinking, one is more likely to say that its past tense is *frinked*; in the context of drinking, one is more likely to say *frank* (Ramscar, 2002). When one hears *The man sawed the woman* while talking about voyeurs, one is more likely to judge it ungrammatical; when talking about stage magic, one is more likely to judge it grammatical.

In summary, a question like “How can one sit comfortably in an office, talk about tigers and galaxies, and be understood?” gets its force by its implicit assertion that meaning is essentially a context-free reference to an arbitrary entity. It is like asking, “How can one fly through the air?” and demanding an abstract answer; there are some contingent answers (“if one is Superman”, “if one is being tossed through the air?”, etc.) but there is no abstract answer, and to attempt to formulate one is to invite confusion. The first response, then, to the tigers-and-galaxies question is to observe that such a situation does not usually happen. We mostly find ourselves in mundane, everyday situations – at the grocery store, chatting with a friend about dinner, talking with a colleague about one’s work, and so on. The second response is to note that there are many kinds of talk about ‘tiger’ (the common noun, a cat with that name, the golf player, a new

product with that name, a metaphorical description, as part of an expletive) and likewise many kinds of talk about ‘galaxy’. So what is the background of the speakers, how did the conversation come about, and what exactly is being said about tigers and galaxies? Such a contextualization is not an evasion of the question of meaning, but is in fact the central message. Context is the ground of meaning.

### Synthesis

In the previous sections, we critiqued some of the shortcomings of the traditional approach to speech. While our critiques were focused on negative points (e.g., that speech is not preceded by a decision, that the word does not have a special status), along the way we sketched out bits and pieces of an alternative approach. Here we pull these bits and pieces together and attempt to lay them out as a coherent framework.

The central tenet of our approach is that speaking and listening are modes of being. What we mean by “mode of being” (equivalently, “mode of activity”) is this: Consider a 6-month old baby crying. It may be crying for one of several reasons – hunger, discomfort, desire to be held – but in no case does the baby think about the reason and then cry; nor is it thinking about the reason *as* it is crying; nor is it conscious of the reason; nor did it plan how to coordinate its breathing and crying. The entire baby – body and brain together – is in the mode-of-crying. Likewise, during speech interactions the speaker is in a mode-of-speaking and the listener is in a mode-of-listening. This view has several implications, which may be summarized as follows.

*Speaking is not two activities (thinking and vocalizing) but only a single activity.* To be in a mode-of-speaking is to be in a certain mode-of-thinking, or more simply, speaking is a form of thinking. We must caution here: there is no such thing as *the* mode-of-speaking or *the* mode-of-thinking. A person can potentially say many things and in many different ways (e.g., mouthing, whispering, shouting). Each of these is a different mode-of-speaking, and there is no one thing (such as jaw, tongue, or vocal cord activity) that is common to all of them. This is not a vague or incomplete definition, but simply reflects the family resemblance structure of our notion of “speaking” (Wittgenstein, 1953, §67).

*When one is in a particular mode-of-being, that mode organizes the activity of the entire person.* Consequently, the mode-of-speaking is not simply the activity of a small set of articulators, it is the activity of the entire person. All aspects of a speaker’s activity – vocalization, posture, breathing, hand gestures, head movements, gaze – are part of the mode-of-

speaking, so they are *all* meaningful.

*The speaker's speech/thought is shaped by the situation that she is in.* For example, if she is with her friend in the park, that puts her in a mode of being that is specific to that particular situation. This mode is shaped by historical context (her history with that friend, that park, and with that friend in that park at that time of day) and immediate context (the conversation thus far, and her own ongoing speech). Her thoughts are shaped by this intersection of historical and immediate context. Put simply, "thought" is context-appropriate skilled activity.

*In saying one thing rather than another, the speaker is not choosing to have that thought; rather, she is choosing to be a certain way.* Usually, this "choosing" just means being sensitive to the situation: if the friend slips and falls, the speaker may be sympathetic (be in a mode-of-sympathy) and commiserate with her friend. If the friend slips on a banana peel and falls, the speaker may be amused but may try to be sympathetic. This conflict (mode-of-amusement vying with mode-of-sympathy) would be reflected in her manner of speech. Sometimes the speaker can decide to resolve a complex issue, and therefore deliberately sustain a particular chain of thought. Here the ability to sustain a chain of thought is a skill (akin to maintaining a complex musical rhythm without being distracted by other rhythms) and, like any skill, it requires training and practice and varies between individuals.

*The listener's sensitivity to the speaker's activity constitutes his mode-of-listening.* The listener, due to his history with the speaker, is sensitized to her modes-of-speaking (her modes-of-thought) that are appropriate for the current situation. In the park example cited above, he is sensitive to how she speaks – how she thinks – when walking in the park. Consequently, he is attentive to her manner of talk, head movement, stance, gestures, gaze, and facial expression. Here "attention" is not a secret process in the mind of the listener, it is what the listener *does* – where and how he looks and listens, how he adjusts his own posture and movement, and so on. Consequently, listening is a skill that needs training and practice. One can listen carefully (attend to all of the speaker's activity, anticipate the progression of her thought) or carelessly (attend to only one or two aspects of the speaker's activity). The careful listener follows along with the speaker, i.e. the listener's mode-of-listening is very finely attuned to the speaker's mode-of-speaking.

*The listener publicly shares in the speaker's thought.* The listener's attunement to the speaker constitutes his understanding of the speaker's thought. When the speaker finishes her turn, the careful listener is able to promptly pick up the "thread" of her thought and elaborate on it. It is rather as if two people take turns to knit a single complex quilt; while one person knits,

the other watches carefully so that he can take over without disruption. Hence, speech is the making-common (the communing) of thought.

Thus far, we have focused on a single speech interaction. But the participants in this one conversation will interact with many other people, who will interact with yet others. Each conversation is a like a tiny fiber, and the language community is a rope made of innumerable many such fibers (Wittgenstein, 1953, §67). To be part of a language community is to be inculcated into its characteristic modes-of-speaking and modes-of-listening, into its characteristic thoughts. When one wants to be fluent in a second language, the difficulty is not simply that of learning the lexicon and grammar of the new language. The main difficulty is that he will have to think and *be* a different way – he will have to become a different person. A Japanese speaker’s difficulty in distinguishing English /r/ from /l/ is not simply an auditory phenomenon, it is an indication of the mismatch or misalignment between characteristic Japanese ways-of-being and characteristic American English ways-of-being. “We may speak several languages, but one of them always remains the one in which we live. In order completely to assimilate a language, it is necessary to make the world which it expresses one’s own, and one never does belong to two worlds at once.” (Merleau-Ponty, 1962, p. 218).

### *Relation to Direct Realism*

As noted earlier, our goal is to show that a detailed conceptual analysis of communication leads naturally to direct realism. Consequently, we deliberately did not invoke direct-realist notions such as affordance and specification. However, at this point it should be clear that our account fits quite naturally within the larger framework of ecological psychology:

- The information in the optic and acoustic arrays specifies the speaker's mode-of-being, i.e. it specifies her thought.
- The listener, by virtue of his shared history with the speaker, can pick up the information specifying the thought. That is, the listener directly perceives the speaker’s thought, without any representational mediation.
- The speaker – while speaking – is attuned to information specifying the listener’s mode-of-being. Hence, the listener's attentiveness regulates the development of the speaker’s own activity.

It is important to note that there is no hierarchy of pickup, i.e. the listener does not first

perceive speech *qua* motor activity and then perceive speech *qua* thought. There is only the perception of thought. In fact, it is almost impossible to perceive speech *qua* motor activity, just as it is almost impossible to perceive light-walking *qua* motor activity (to perceive order in the light patterns is to see purposive activity such as climbing stairs or sitting down or walking). Even if the listener is not familiar with the speaker's language or dialect, he still perceives the speaker's thoughts, though there would be a greater chance of misunderstanding.

Several of the above ideas have been articulated previously. For example, John Dewey observed that “the heart of language is not ‘expression’ of something antecedent, much less expression of antecedent thought. It is communication; the establishment of cooperation in an activity in which there are partners, and in which the activity of each is modified and regulated by partnership.” (1925, p. 179). Similar thoughts are found in Jenkins’ (1977) advocacy of “contextualism”, and in Verbrugge's (1985) proposal that speech is the direct perception of communicative events. In particular, Verbrugge’s observation that “symbols emerge as *constituents* of natural events” (p. 183) dovetails nicely with our own view.

We should also point out a key difference between his view and ours. Verbrugge notes that the concept of affordance may not be applicable to social action, because “these events have no obvious ‘dual’ in action at all” (p. 189). That is, the same utterance may be understood differently by different listeners and worse, the listeners’ responses affect the development of the thought. We suggest that the concern can be addressed by enlarging the notion of affordance. An affordance is a relation between a molar property of the environment and the complementary motor capability of an organism (Chemero, 2003). Typically, the molar property and the motor capability are treated as givens. For example, a cave affords shelter for a rat; this affordance is not changed when a rat approaches it, nor is it significantly different across rats of the same species. In such a case, a researcher can try to identify the optic invariant that specifies “shelter” for a rat. But a social situation requires a more general approach: the molar property does change, and the individual history of the organism is profoundly important. It is therefore not possible, even in principle, to identify *the* optic/acoustic invariant that specifies a certain social affordance. Such an invariant can be identified only for specific, highly-contextualized cases.

Our view also clarifies the theoretical status of the direct perception of articulatory events (DPAE). DPAE implies that speech is composed of articulatory gestures, but we have argued earlier, such compositionality is untenable. In everyday speech the listener directly perceives the thought and there is no need to perceive the articulatory gesture. However, the distinction between thought and gesture is not a sharp one, and some situations do require the listener to attend to the articulation. For example, one may listen in order to imitate a language teacher;

repeat a pledge; learn an alphabet; learn an unfamiliar name; learn a certain accent; produce a phonetic transcription, and so on. Presumably, DPAE is applicable in such situations. But two caveats need to be noted. First, the cited situations are rather specialized. A listener is “by default” sensitive to communicative events and needs to learn to attend to articulatory events, just as the artist needs to learn to attend to shape and shading rather than to the scene. Second, repeating a pledge calls for a different kind of listening than producing a phonetic transcription. There is no such thing as a “pure” DPAE, but rather there are many varieties of situation-specific DPAEs, each of which has to be learnt. Thus the larger point is (again) the importance of situation and context.

### *Experimental Approaches*

It is easy to be a critic. All one needs to do is think very hard about any complex aspect of the world and it quickly becomes apparent why this or that approach to its study is defective in some way. It is rather more difficult to suggest how we can, in practice, do better (Lewontin, 2000, p. 109).

The above framework of speech and language is based on a conceptual analysis, but such an analysis does not point the way to a research program. If speech is thought, and thought is inseparable from the situation, where do we start studying it? How can the theory be elaborated beyond “Everything is connected to everything else”? One avenue of study is the importance of systematic fine phonetic detail in speech. In recent years there has been renewed interest in the variety of such detail (e.g., Hawkins, 2005) and how it is influenced by social interaction (Krauss & Pardo, in press). This interest is driven by the insight that speech perception and production are closely tied to the circumstances of the utterance, and consequently it aligns closely with the theoretical perspective being proposed here. In addition, we suggest a few more avenues of exploration.

- Speaking is not simply a vocalization, it is a mode of being. A listener is sensitive to all the activity of the speaker – facial expression, direction of gaze, head and torso movement, gestures, stance, etc. – and disturbance of any these activities affects the perception. For example, if listeners are played a /ba/ sound while shown the lip movement for /ga/, they are more likely to hear /da/ (McGurk & McDonald, 1976). Relatedly, listeners are better able to identify syllables if the speech is paired with a video of natural head movement (Munhall et al., 2004). We should expect similar effects if the facial expression and gestures of the speaker are similarly manipulated. For example, speech perception should be more error-prone if (in the appropriate context) the speaker says “down” while gesturing up, “hello”

without initiating eye contact, or “no” while nodding (see McNeill, 1987, for a more comprehensive treatment of gesture and speech).

- Speech perception is influenced by the speaker’s and listener’s relation to their surroundings. For example, the perception of “help me up” may be facilitated if the speaker is below the eye level of the listener and extending an arm, versus when the speaker is standing or not extending an arm. More generally, speech perception should be impaired if the speaker exhibits irregular or situationally-aberrant eye, head or body movement while speaking (such as pointing or looking at a cat while saying “dog”).
- The listener’s activity is essential to speech perception. One kind of activity is eye movement. Altmann (2004) showed subjects a display of several objects, blanked the display, and played a sentence referring to those objects; while hearing the sentence, the subjects directed their gaze toward the appropriate regions of the blank display. Altmann interpreted this as evidence for a “mental record”, but we suggest that the eye movements are constitutive of the speech perception. For example, if the eye activity is disrupted while the sentence is being heard – by showing random activity, by showing a new and unrelated scene, or by stabilizing the retinal image (Blakemore et al. 1971) – then the speech perception should be more likely to be impaired. Another kind of activity is body movement. If the listener is performing some motor task while listening, then speech perception should be worse if the task is unrelated to the utterance- and situation-context. Such impairment has been shown, for example, when subjects are asked to tap while perceiving and recalling visually presented letters (Larsen & Baddeley, 2003).
- “The situation” is an ensemble of contextually-appropriate relational invariants. If, in the appropriate context, the listener is sensitive to the speaker nodding while saying “yes”, his speech perception would not be impaired if the head movement is replaced by the optic invariant for “nodding”. Such robustness has been demonstrated when the acoustic signal is replaced with a sine-wave analogue, and when the speaker’s face is replaced with a point-light analogue (Rosenblum, 2004). We predict that speech perception would also be robust when speakers’ gestures are replaced with point-light analogues, or when parts of the situation are replaced by optic or acoustic invariants.
- To learn a new linguistic expression is to learn a new kind of situation. Word learning has been shown to be specific to speaker identity (Nygaard & Pisoni, 1998), speech rate (Bradlow et al., 1999) and emotional tone (Mullennix et al., 2002), but we suggest that these are instances of an overall situation-specificity. We can expect learning to be specific to

manner of speech (whispered, shouted, slurred), mode of interaction (face-to-face, over the phone while watching a blank screen, over the phone while watching a video), topic of conversation, overall setting, and so on.

The central theme of the above proposals is that speech is situation-specific. But a 'situation' gains its significance only through its relations to innumerable other situations. Therefore the essence of speech perception is not an activity common to the various situations, but rather the essence is the variety itself. We see the scientific task here as mapping the distribution and morphological variety of speech-situations; the relations within a given situation (such as the coordination of the speaker's vocal and body activity and the interlocking between the speaker's and listener's activities); the differentiation of existing situations into new varieties; the adaptation of speakers and listeners to changing situations (e.g., Sancier & Fowler, 1997), and so on. It is beyond the scope of this paper to develop these ideas, but we want to highlight a connection to existing theory.

In the 1960s, the psychologist Roger Barker noted that "The descriptive, natural [historical], ecological phase of investigation has had a minor place in psychology, and this has seriously limited the science" (R. Barker, cited in Heft, 2001, p. 245). He proposed that the activity of individuals is largely shaped by their "behavior setting" – such as a lecture, a baseball game, or a worship service – and set out a program of *ecobehavioral science* to describe and catalog various behavior settings (Barker, 1978). This program is closely related to direct realism (Heft, 2001), and moreover fits naturally into the formalism of coordinative structures (Kelso & Tuller, 1981) – "Behavior settings are best conceptualized as dynamic systems operating as contexts of constraint. They are time-dependent phenomena, with their boundaries established and maintained by the coming together of particular behavior-milieu components ... [They are also] quasi-stable systems in that they can withstand minor perturbations, often in novel ways." (Heft, 2001, p. 321). In general, Barker's 'behavior setting' is similar to what we have called a 'situation', though we would emphasize that a situation can be more than a geographic location. When we are conversing over a phone, we are socially *with* the other person – we feel their presence – so the phone conversation too is a kind of situation.

### Reflections

The primary goal of this paper was to formulate a direct realist view of language. We hope we have demonstrated that language is not merely compatible with direct realism, but rather that a proper account of language requires a direct realist approach. Central to our demonstration is

the use of phenomenology and ordinary-language analysis, and it is instructive to examine what these approaches allowed us to do.

Theorists advocating the ecological approach often stress that it is a *logical error* to treat perception as involving the reconstruction of the world from sensory stimuli (e.g., Shaw & Bransford, 1977; Reed, 1996). While the argument is valid, it typically does not get traction with many psychologists (e.g., Diehl, Walsh & Kluender, 1991). It is unlikely that these researchers are simply blind to the logical error, so what is the reason for the disagreement? We suggest that one factor is a focus on certain canonical examples. For instance, a listener *can* transcribe sentences presented over headphones; a listener *can* learn the meaning of a word through ostensive definition; and speakers *can* sometimes lie. Researchers focused on these examples will take it as axiomatic that all speech is akin to a transcription, that all words are learnt through ostensive definition, and that speakers' intentions are always uncertain. A theoretical critique of such a position often has no force because it leaves the underlying examples – and the theoretical language that reinforces those examples – in place. To be effective, a critique needs to weaken the grip of the canonical examples and bring out the diversity of the phenomenon under study. Such a critique is precisely what is provided by a phenomenological and ordinary-language analysis. The analysis also helps to characterize the natural “joints” of a phenomenon, and can therefore guide the development of ecological approaches in other domains of cognition.

## References

- Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: the 'blank screen' paradigm. *Cognition*, 93(2), B79-B87.
- Austin, J. L. (1962). *Sense and Sensibilia*. Oxford: Oxford University Press.
- Barker, R. G. (1978). *Habitats, Environments, and Human Behavior*. San Francisco: Jossey-Bass Publishers.
- Bennett, M. R., & Hacker, P. M. S. (2003). *Philosophical Foundations of Neuroscience*. Oxford, UK: Blackwell.
- Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience* (pp. 171-204).
- Blakemore, C., Muncey, J. P. J., & Ridley, R. M. (1971). Perceptual fading of a stabilized cortical image. *Nature*, 233, 204-205.
- Bloomfield, L. (1933). *Language*. New York: Holt, Rinehart and Winston.
- Bradlow, A., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics*, 61(2), 206-219.
- Bybee, J. (2001). *Phonology and language use*. Cambridge, UK: Cambridge University Press.
- Chafe, W. L. (1982). Integration and involvement in speaking, writing and oral literature. In D. Tannen (Ed.), *Spoken and Written Language* (pp. 35-52). Norwood, NJ: Ablex Publishing.
- Chambers, J. K. (1995). *Sociolinguistic theory*. Oxford, UK: Blackwell.
- Chemero, A. (2003). An outline of a theory of affordances. *Ecological Psychology*, 15(2), 181-195.
- Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.
- Dewey, J. (1925/1958). *Experience and Nature*. New York: Dover Publications.
- Diehl, R. L., Walsh, M. A., & Kluender, K. R. (1991). On the interpretability of speech/nonspeech comparisons: A reply to Fowler. *Journal of the Acoustical Society of America*, 89(6), 2905-2909.
- Dreyfus, H. L. (1991). *Being-in-the-World*. Cambridge, MA: MIT Press.
- Dreyfus, H. L. (1992). *What computers still can't do*. Cambridge, MA: MIT Press.
- Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14, 3-28.
- Fowler, C. A. (1991). Auditory perception is not special: We see the world, we feel the world, we hear the world. *Journal of the Acoustical Society of America*, 89(6), 2910-2915.
- Fromkin, V., & Rodman, R. (1978). *An introduction to language* (2nd ed.). New York, NY: Holt, Rinehart and Winston.

- Gibbs, R. W., Jr. (1994). *The poetics of mind*. Cambridge, UK: Cambridge University Press.
- Glotzbach, P. A., & Heft, H. (1982). Ecological and phenomenological contributions to the philosophy of perception. *Nous*, 16, 108-121.
- Harris, J. (1994). *English sound structure*. Oxford, UK: Blackwell.
- Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31(3-4), 373-405.
- Heft, H. (2001). *Ecological psychology in context*. Mahwah, NJ: Lawrence Erlbaum.
- Heidegger, M. (1927/1977). *Being and Time* (translation, Trans.). New York: Harper Collins.
- Jackendoff, R. (2002). *Foundations of language*. Oxford, UK: Oxford University Press.
- Jenkins, J. J. (1977). Remember that old theory of memory? Well, forget it! In R. Shaw & J. Bransford (Eds.), *Perceiving, Acting and Knowing* (pp. 412-430). Hillsdale, NJ: Lawrence Erlbaum.
- Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145-165). San Diego, CA: Academic Press.
- Kelso, J. A. S., & Tuller, B. (1981). Toward a theory of apractic syndromes. *Brain and Language*, 12, 224-245.
- Krauss, R. M., & Pardo, J. S. (in press). Speaker perception and social behavior: Bridging social psychology and speech science. In P. A. M. v. Lange (Ed.), *Bridging Social Psychology*. Hillsdale, NJ: Lawrence Erlbaum.
- Larsen, J. D., & Baddeley, A. (2003). Disruption of verbal STM by irrelevant speech, articulatory suppression, and manual tapping: Do they have a common source? *The Quarterly Journal of Experimental Psychology*, 56A(8), 1249-1268.
- Lewontin, R. (2000). *The triple helix: Gene, organism and environment*. Cambridge, MA: Harvard University Press.
- Local, J. (2002). *Variable domains and variable relevance: Interpreting phonetic exponents*. Paper presented at the Temporal Integration in the Perception of Speech (TIPS) Conference, Aix-en-Provence, April 8-10, 2002.
- Lord, A. B. (1982). Oral Poetry in Yugoslavia. In U. Neisser (Ed.), *Memory Observed* (pp. 243-257). New York: W. H. Freeman and Company.
- Mace, W. M. (1977). Ask not what's inside your head, but what your head's inside of. In R. Shaw & J. Bransford (Eds.), *Perceiving, Acting and Knowing* (pp. 43-66). Hillsdale, NJ: Lawrence Erlbaum.
- Maratsos, M. F. (1977). Disorganization in thought and word. In R. Shaw & J. Bransford (Eds.), *Perceiving, Acting and Knowing* (pp. 347-363). Hillsdale, NJ: Lawrence Erlbaum.
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.
- McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences*, 6(11), 465-472.

- McDowell, J. (1998). *Mind, Value and Reality*. Cambridge, MA: Harvard University Press.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.
- McNeill, D. (1987). *Psycholinguistics: A new approach*. New York: Harper & Row.
- Merleau-Ponty, M. (1962). *Phenomenology of Perception* (Colin Smith, Trans.) (Vol. Original work published 1945). London: Routledge.
- Mullennix, J. W., Bihon, T., Bricklemeyer, J., Gaston, J., & Keener, J. M. (2002). Effects of variation in emotional tone of voice on speech perception. *Language and Speech*, *45*(3), 255-283.
- Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, *15*(2), 133-137.
- Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355-376.
- O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*, 939-1031.
- Pessoa, L., Thompson, E., & Noë, A. (1998). Finding out about filling in: A guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences*, *21*(6), 723-748.
- Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology VII* (pp. 101-140). Berlin: Mouton de Gruyter.
- Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*(11), 456-463.
- Ramscar, M. (2002). The role of meaning in inflection: Why the past tense does not require a rule. *Cognitive Psychology*, *45*, 45-94.
- Reed, E. S. (1996). *Encountering the World: Toward an Ecological Psychology*. New York: Oxford University Press.
- Richardson, M. J., Marsh, K. L., & Schmidt, R. C. (2005). Effects of visual and verbal interaction on unintentional interpersonal coordination. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(1), 62-79.
- Rosenblum, L. D. (2004). Perceiving articulatory events. In J. G. Neuhoff (Ed.), *Ecological Psychoacoustics* (pp. 219-248). Academic Press.
- Ryle, G. (1949). *The concept of mind*. Chicago: The University of Chicago Press.
- Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, *25*(4), 421-436.
- Sartre, P. (1943/1993). *Being and Nothingness* (Translated.). New York: Simon and Schuster.
- Shaw, R., & Bransford, J. (1977). Psychological approaches to the problem of knowing. In R. Shaw & J. Bransford (Eds.), *Perceiving, Acting and Knowing* (pp. 1-39). Hillsdale, NJ: Lawrence Erlbaum.

- Suzuki, T. (2001). *Words in context: A Japanese perspective on language and culture* (Akira Miura, Trans.). New York: Kodansha International.
- Tannen, D. (1984). *Conversational style: Analyzing talk among friends*. Norwood, NJ: Ablex Publishing Corporation.
- Trask, R. L. (1996). *A dictionary of phonetics and phonology*. New York: Routledge.
- Varela, F. J., Thompson, E., & Rosch, E. (1991). *The Embodied Mind*. Cambridge, MA: MIT Press.
- Verbrugge, R. R. (1985). Language and event perception: Steps towards a synthesis. In J. William H. Warren & R. E. Shaw (Eds.), *Persistence and Change* (pp. 157-194). Hillsdale, NJ: Lawrence Erlbaum.
- Wittgenstein, L. (1953). *Philosophical Investigations* (G.E.M. Anscombe, Trans.) (3rd ed.). Upper Saddle River, NJ: Prentice Hall.