

# Language & Cognition

2000–2001

*University Seminar #681*

Columbia University  
New York, New York





# *Language & Cognition*

---

What can the study of language contribute to our understanding of human nature? This question motivates research spanning many intellectual constituencies, for its range exceeds the scope of any one of the core disciplines. The technical study of language has developed across anthropology, electrical engineering, linguistics, neurology, philosophy, psychology, and sociology, and influential research of the recent era of cognitive science has occurred when disciplinary boundaries were transcended. The seminar is a forum for convening this research community of broadly differing expertise, within and beyond the University. As a meeting ground for regular discussion of current events and fundamental questions, the University Seminar on Language and Cognition will direct its focus to the latest breakthroughs and the developing concerns of the scientific community studying language.

University Seminar #681, Founded: 2000

## SEMINAR ADMINISTRATION

CHAIR: Robert E. Remez  
Ann Whitney Olin Professor  
Department of Psychology, Barnard College  
(212) 854-4247  
remez@columbia.edu

RAPPORTEUR: Ezequiel Morsella  
Doctoral candidate in Psychology, Columbia University  
(212) 854-7033  
morsella@psych.columbia.edu

WEBPAGE: <http://www.columbia.edu/~remez/langcog.html>



## *Table of Contents*

1. Why speakers imitate; how they communicate CAROL A. FOWLER .....	7
2. The perceptual organization of speech ROBERT E. REMEZ .....	19
3. The cognitive consequences of linguistic change WILLIAM LABOV .....	29
4. A deficit in visual location perception: Implications for spatial representation and reading MICHAEL McCLOSKEY .....	37
5. Infants' segmentation and recognition of words PETER W. JUSZYK .....	43
6. Universal and evolutionary aspects of cross-language color naming PAUL KAY .....	51
7. Language acquisition and intentionality: The essential tension between engagement and effort LOIS BLOOM .....	57



28 SEPTEMBER 2000

---

## Why speakers imitate; how they communicate

Carol A. Fowler  
*President, Haskins Laboratories,  
New Haven, Connecticut*

Chomsky (*Knowledge of Language*, 1986) has argued that externalized (E-) language, or public language use, corresponds to “no real world object” and cannot be studied scientifically. In these respects, E-language contrasts with internalized (I-) language, or knowledge of language in the human mind. I will suggest that the arguments against studying E-language are mistaken and have hindered the development of our understanding language by focusing researchers’ attention almost exclusively on the minds of individual language users. Understanding language processing and language competence is important, but so is our understanding of public language use for an appreciation of the fundamentally social and cooperative function of language in human life. I will show how a focus on public language use, and specifically here on cross-person linguistic imitation, can illuminate our understanding of communication by language at the phonological level of description and our understanding of the role of language in human life more generally.

In 1957, B. F. Skinner published his book *Verbal Behavior* in which he attempted to show that the laws of learning that he considered to be universal, although they had been largely uncovered in the rat and pigeon laboratories, were powerful enough to explain that most complex of human behaviors: verbal behavior. In 1959, Noam Chomsky published a scathing review of that book. He argued, for example, that the concept of reinforcement history, which can be operationalized in the rat lab because the experimenter provides the relatively important history, became circular in the context of the book, because Skinner was reduced to observing what people said in a particular context and then guessed what their reinforcement history must have been—and, of course, if you do that, you can’t lose. Chomsky also argued that there can be no scientific setting of public language use, that to study language must be to study the mind. His review of the book was extremely impactful, and, in particular, in the field of psychology it helped promote the near eclipse of behaviorist approaches to understanding verbal behavior and helped the rapid development of the field of cognitive psychology, and within that, psycholinguistics. And if you look at what goes on in the field of psycholinguistics, you see an impact of his thinking and arguments there, too.

In psycholinguistics, largely we study what goes on inside the head of the language user when they perceive speech, comprehend language, plan an utterance, and when they implement that plan as vocal tract activity. Or, we study what goes on inside the head of a child as the child becomes a

competent user of his or her language. There is very little study of public language use.

Let me say that there is some study of public language use: Robert M. Krauss studies it, and Herbert S. Clark at Stanford studies it, and some of their students do; but it's not a very well developed field. What I'm trying to argue today is that we need an understanding of public language use, and I'll try to show how focusing attention on language use in public can be illuminating in two domains: one, the domain in which I spent most of my career doing research—the study of phonological production and perception. The other domain is the use of language in cooperative, social activity.

Why, in Chomsky's view, can there be no scientific understanding of public language use? Well, here is what he wrote in his 1986 book *Knowledge of Language*. He makes the distinction between what he calls *E-Language*, or *externalized language* that is public language use and *I-Language*, which is language in the mind, or *internalized language*. Chomsky says,

“E-language, that was the object of study in behavioral psychology is now regarded as an epiphenomena at best. Languages in this *E-Language* sense are not real world objects, but are artificial, somewhat arbitrary, and perhaps not very interesting constructs. Universal grammar and theories of *I-Language*, universal and particular grammars, are on a par with scientific theories of other domains. Theories of *E-Language*, if sensible at all, have some different and more obscure status, because there is no corresponding real world object.”

I'm not sure I know what Chomsky meant by what he wrote there. My first thought was—in trying to understand why he would say that public language use is not a real world object—that in defining *E-Language* he invokes the definition of language of the structural linguist Leonard Bloomfield, a sort of behaviorist linguist. Bloomfield defined language as something like this: “the set of all utterances that members of a language community either have uttered or might utter.” Certainly if you defined language that way it doesn't correspond to a real world object—there are lots of utterances that haven't been said yet, and the particular subset that has been said is not a natural subset of all possible utterances. But that's not what Chomsky meant, because he goes on to say, “the concept of E-language, however construed, appears to have no significance.” What he might mean relates to his distinction between competence and performance. As I stand here speaking to you, what is influencing what I say is only partly my language history; in addition, there is how much attention I'm putting on to what I'm saying, how much I am speaking while monitoring my train of thought, etc. There are all kinds of non-linguistic influences on my public language performance, and perhaps that what Chomsky meant when he said that *E-language* corresponds to no real world object. I'm going to suggest, though, that we have to try to develop a scientific understanding of public language use, if we are going to understand human life, because the public use of language is what we do, and it is a very significant part of our life.

In the domain of speech production and perception—let me begin there—we know that spoken language use was important in our evolutionary

history. We know that it was important because there are adaptations of the brain and adaptations of the vocal tract to be used for spoken language.

*[Professor Fowler shows a diagram of human and nonhuman primate vocal tracts.  
Taken from Phillip Lieberman's Uniquely Human.]*

Lieberman points out a marked difference between the human vocal tract and the nonhuman primate vocal tract. The velum and epiglottis in the primate are in contact with each other; they can be used to form a seal. But in the human they are widely separated. This is because, in the human, the larynx has descended in the throat. And that has two consequences according to Lieberman: 1) humans can produce sounds that nonhuman primates can't produce, and 2) humans are more likely than other primates to choke on their food, and sometimes to die as a result. This, according to Lieberman, is an adaptation for speech that would otherwise be maladaptive. So, spoken language use was important in our evolutionary history and was important enough that the physical adaptations of the brain and vocal tract have taken place. We might expect, then, if public language use was important in our evolutionary history, that language itself was adapted for public language use.

What would we count as an adaptation of language itself? Well, one obvious property, which is uncontroversial, is to have language form. If I am going to try and communicate with someone out in the world, I have to behave in public in a perceivable way that counts as behaving linguistically. Language forms let us do that. Practically, if we only used language in our heads we wouldn't need the language forms; we would only need the meanings. But I think we could say more than that. We use speech in our lives in a variety of ways to accomplish a variety of purposes; so *speech act theory* can provide taxonomy of the ways we produce speech. But there is a bottom-line requirement for a speech act to be effective: the listener has to hear what the talker has to say. If I'm standing in a room near a door, and someone else is standing near an open window, and I say, "I'm cold," you might interpret that as a very polite way of my asking you to close the window. In order for you to interpret it you have to have a fairly sophisticated understanding of indirect speech acts and how they work, but you also must hear what I said. If you didn't hear me say "I'm cold," that speech act has no way of getting through. What's required for spoken communication to have any possibility of being effective is that, prototypically, there needs to be a relation of sufficient equivalence between messages sent and messages received, and I'm going to call that a *parity requirement*.

I am going to suggest that, if spoken communication was important in our evolutionary history, we might expect language to have *parity-fostering characteristics*. What might those be? I'm going to propose two of them. One parity-fostering characteristic would be if the language forms we use were somehow isomorphic with the public actions we engage in when we talk. That's the most transparent way in which I can convey my linguistic messages, that is, if my language actions are language forms themselves. Another parity-fostering characteristic is the preservation of the forms throughout the communicative exchange, that is, the speaker is going to intend to produce the

language forms and is in fact going to produce them by an acoustic signal that conveys the language form to the listener.

In the field of psycholinguistics, do we find in speech production and perception a language that has these parity fostering characteristics? I think we don't. If you ask a speech scientist, someone who studies production or perception, about phonological language forms, he or she will say that they are abstract categories in the mind of language users. They have the property that they are discrete units—e.g., consonants and vowels. When they are described as featural attributes, they are static, and they are context-free: a /d/ is a /d/ no matter where it occurs. If we now compare those attributes of phonological forms, as people know them, to what goes on in the vocal tract, we see something quite different. When we speak, we co-articulate, which means that we temporally overlap gestures corresponding to the consonants and vowels of the successive segments of words. What we see going on in the vocal tract is nothing discrete, nothing static, and movements that are highly context sensitive. There seems to be a lack of correspondence between what goes on in the vocal tract and what the speaker intended to convey to the listener.

There is a general view that coarticulation is a destructive process on the essential properties of phonological segments. The most vivid metaphor for coarticulation, and perhaps an exaggeration, as most people would agree, was put forth by Charles Hockett. Hockett said that the effect of coarticulation on phonological segments is like the effect on an array of brightly colored Easter eggs, being consonants and vowels, if you were to pass them through a wringer. By doing this, you would lose discreteness, you would lose context-freeness. If that's the case, if coarticulation is like smashing Easter eggs, then the acoustic signal could not be a transparent window on the phonological segments that the talker is trying to convey, and, in fact, in the field of speech perception, for the most part, theory suggests that perception is a reconstructive process: you take those bits and pieces of egg yoke and shells and try to get those eggs back in shape. If you remember the difficulties that all king's horses and all the king's men had, you'll realize that that is not an easy task, that perception has to be reconstructive.

This, I think, is a canonical way of looking at speech production and speech perception. Do we *have* to look at it this way? I don't think so. I think the reason why this viewpoint has not been challenged is because we haven't paid very much attention to what people using language try to do. People who study speech production tend not to study speech perception, and such scientists don't care if the articulatory gestures of the speaker are communicating effectively to a listener. People who study speech perception tend only to study speech perception and not how speech perception plays a role in linguistic communication. For the last fifteen or twenty years, my colleagues and I at Haskins Laboratories have been trying to replace that scenario with one in which language forms—phonological forms—are the actions of the vocal that we use when we speak. So this is the first parity-fostering characteristic: language forms *are* what we do when we speak. We call these *gestures*. Gestures are linguistically significant actions of the vocal tract. Characteristically they create and release constrictions. So, for example, when I shut my lip to say /b/, /p/ or /m/, that is a gesture. It is a coordinated activity

of my jaw, upper lip, and lower lip. When you look at vocal tract activity in this more macroscopic way, not looking just at the jaw, lips, or tongue, but rather at the whole gesture, coarticulation does not look like a destructive process. I always close my lips shut when I pronounce /b/, yet /b/ is considered to be a low-resistance consonant in that it resists coarticulatory encroachment from its neighbors very little. There is a lot of coarticulation that goes on because of neighboring vowels, yet we always manage to get our lips shut. At the macroscopic level, we don't need to see coarticulation as destructive.

Gestures are public actions. What is in my head are not gestures but what I know about gestures. The phonological forms are not in the head, but what I know about them are, in the same way that what I know about elephants is in my head. Elephants don't reside there, just my knowledge of them. In this scenario, then, speakers plan to produce sequences of words which are composed of gestures and implement those gestures non-destructively as vocal tract activity. The gestures out in public are structuring the acoustic signal which provides information about the signal, and, we claim, that's what listeners perceive. There are a number of aspects of this scenario which are controversial. There is a theory of phonology developed by my colleagues Louis Goldstein and Kathe Broman on articulatory phonology. This theory attempts to build a linguistic phonology out of these gestures, which is controversial. My claim that coarticulation is not destructive is controversial. The thing that I'm going to focus on that is controversial is that listeners of speech perceive gestures.

When I gave a practice version of this talk someone said, "You should say right up front that you are not speaking about watching the movements of people's mouths in perceiving speech." Of course people perceive speech by telephone where visual information about the mouth movements are lacking. I am saying that the acoustic signal provides information about the gesture, because the gestures were the cause of the structure of the acoustic signal, in the same way that the structure of the light that strikes my eye provides information about the properties of the objects it reflected off from. I don't see reflected light; I don't hear acoustic signals. I hear what they tell me about their distal cause, their cause in the environment. I think that there are many experiments that have demonstrated that listeners of speech perceive the significant actions of the vocal tract. I will not rehearse them all here for you; I am just going to talk about the oldest and newest findings that I know.

In 1950's Alvin Liberman and his colleagues were trying to track down the so-called acoustic cues for consonants and vowels. They were trying to discover what physical acoustic structure was required in order for listeners to identify a consonant as a /b/, for example. There are consonants known as stops in most languages, produced by stopping the air flow in the vocal tract temporarily. In English, /b/, /d/, /g/, /p/, /t/, and /k/ are the stop consonants. Liberman's research uncovered that these consonants are cued in two fundamental ways. We make the constriction of the lips for a /b/ or a /p/, and release the constriction fairly explosively and release a little bit of energy known as a burst.

*[Prof. Fowler shows a spectrographic representation of a burst.]*

The other thing that happens is that as the constriction opens up to the vocal tract shape of the vowel, we begin to see *formants*, which are resonant characteristics of the vocal tract. Formants are acoustic energy at certain resonances, and as the vocal tract opens, we see *formant transitions*, because the acoustic cavity is rapidly changing its size and shape. So the cues are the burst of energy and the formant transition. Liberman's work showed that in synthetic speech, either cue alone is sufficient to lead to the percept of a stop-consonant or a vowel. Two of his important findings, findings which made him a motor theorist, were obtained with these stop-consonants. Liberman's research showed that it is the transition of the upper formant, the second formant, that provides information about the consonant at the beginning of the syllable. So it is the rapid falling energy in /du/ and rapid rising energy in /di/ that underlie the perception of the consonant /d/ and the vowels. But what is surprising is how different these formant transitions are depending on the vowel. If you isolate the consonants from their vowels, they sound as you would expect them to by looking at the spectrograph: they sound like chirps. Why do we hear /d/ in both of these syllables (/di/ and /du/)? Well, Liberman realized that, though their respective spectra are very different, there is something which is the same about these two syllables: how we produce them, that is, the articulatory gestures. The reason the formant transitions are different, in the context of the different vowels and coarticulation, is because while we use the tongue tip in forming the /d/ the rest of the tongue is conforming itself for the vowel. Therefore, the resonance characteristics of vocal tract activity is different, and the formant transitions are different. So we have one articulation—the tongue-tip gesture—that because of coarticulation gives rise to different acoustic signals. The listener appears to track articulation. Hence, Liberman said, "when articulation and acoustics go their separate ways, which way does perception go? It goes with articulation."

Here is a complimentary example. Now we are ignoring the formant transitions, but we are using the burst of energy to cue a consonant. If we take the burst of energy at 1440 Hz, and we place it in front of an /a/ or /i/ vowel, most listeners report hearing a /pi/. Take that same burst of energy, and place it in front of an /a/ vowel, and most listeners report hearing /kɑ/. Here we have one acoustic cue—the burst of energy—that, because of coarticulation, had to be produced in two different ways for the different vowels—by the lips in front of the /i/ vowel and by the tongue body going against the soft palate in the context of the /a/ vowel. Again, the listener is hearing the articulation. These are the findings that led Liberman to develop his motor theories of speech perception. Those are the oldest findings I know.

In 1980, Bob Porter and co-authors published a couple of papers in which they recorded a remarkable finding in a simple-choice reaction-time experiment. In a simple reaction-time experiment, a listener presses a response button after hearing a tone. In a choice reaction-time experiment, the subject must press one of two buttons after he or she makes a decision. For example, for a high tone they would hit one button; and for a low-tone, they would hit another button. Canonically, in studies like this, there is about a 100 – 150 ms

difference when choice is introduced. Porter reported that that difference can disappear when the choice tasks are appropriately chosen speech tasks.

*[Professor Fowler shows data from a recent experiment of hers on the subject, in which the findings were replicated. The experiment demonstrates that, when the choice task is a speech perception task—detecting consonants, the difference between simple and choice reaction times was 26 ms, which, although a statistically significant difference, was substantially less than the 100 – 150 ms difference found in non-speech decision tasks.]*

Why are people so fast at this decision task? If you think that people perceive acoustic speech signals, there is no answer to this question. If you perceive gestures, however, you perceive that the model shut her lips. What you perceive are instructions for your own response in the choice task. If we perceive gestures, it is understandable that the element of choice has diminished in magnitude relative to that of a standard choice task. We also took a look at reaction times in a simple reaction task, because we thought, if a subject is perceiving the gestures and those gestures are acting as a goad, a subject must be faster, even in the simple task, when the model's consonant matches the one they must produce. Although there are very few data (only six subjects), in one of the two observations, the results support this hypothesis. Response times are shorter when there is a match between what the model produces and what the subject must articulate. Finally we took our model's utterances and lengthened the voice-onset time (VOT) of half of them. The VOT is the time from release of the consonant to when the voicing for the vowel turns on. We doubled that interval, and ran another set of subject ( $n = 12$ ) through the choice reaction time task. Our question was, will subjects imitate by producing longer VOT responses? The answer is yes. So we imitate one another very quickly and very easily.

I conclude that a focus on public language use suggests that languages ought to be parity-fostering. This suggests that talkers should produce and listeners should perceive public language forms, which I identified as gestures. The finding that imitation is fast and easy supports this claim.

Now I want shift gears and focus on the production of language in cooperative settings. I will shift my attention away from findings suggesting that imitation is easy and fast and look at the observation that imitation is dispositional. When I was at Dartmouth College in the 1980's, some of my colleagues did an imitation study. They had videotaped some footage of President Reagan on the campaign trail that they showed to two groups of subjects. One group of subjects—an easy group to find at Dartmouth—was made of fans of Reagan; the other group consisted of people who did not like Reagan. They put surface electrodes on the muscles that we use to smile and on the muscles that we use when we frown, and they recorded the muscular activity of the subjects as they watched these silent video clips. They found that when Reagan smiled it would activate the smiling muscles in subjects, and that when he frowned it would activate the frowning muscles in the subject, regardless of whether or not the subject had a favorable view of the candidate. The subjects in the two groups were indistinguishable in respect to their imitation of Reagan. We also know that humans imitate from a very young age. Infants imitate facial gestures—e.g., sticking out the tongue—as early as 32

hours after birth. This is an astonishing accomplishment considering that the infant has never seen his own tongue. How does the infant know what to do? If we ask why people are disposed to imitate, there are a variety of answers. One of them is that through imitation, youngsters become members of their cultural group. But I don't think that that is the only answer, because adults tend to imitate too. Why?

In his book *Using Language*, Herbert Clark points out that prototypical public language use occurs in the context of cooperative, coordinated activity which Clark calls *joint activity*. For example, people use language to communicate while cooking together or while purchasing something at the pharmacy. In most of these cases, it is astonishing how relatively little speech had to occur for the interaction to take place. The speech that *did* take place was just that which was required to push that joint activity forward and to be polite with each other. My thought about imitation is that, as social beings, we often coordinate the self with another person or with other people. In the course of doing that we entrain to one another in ways that I will describe. I think that imitation is the most primitive function by which we coordinate with each other. There are a lot of studies that show that when people are interacting cooperatively they tend to imitate or entrain with one another. For example, during cooperation, people converge in dialect, speaking rate, vocal intensity, and their rate and duration of pausing. When listeners find a speaker engaging, they tend to imitate that speaker's skeletal posture. There is also *interactional synchrony*—speakers are alleged to move in time with the listener's rhythm, and vice versa.

There are many findings of this sort in the literature, so I am inclined to believe in them. However, many of them use highly subjective measures. For example, in the interactional synchrony literature, a lot of the video coding is done by hand. With some students, I have attempted to devise more objective procedures to study cooperative language use. I will describe two studies. The first was done by Jennifer Pardo at Yale, as part of her dissertation research. Jennifer used a cooperative map task in which people work as partners.

*[Professor Fowler shows a sample map and goes on to describe the experiment. In this task, there is an instruction-giver and an instruction-receiver. Sometimes the giver and receiver have different maps. The object is for the receiver to reach a destination on the map, though his or her map is somewhat different from the instruction giver's map. Pardo wanted to look at speech imitation, by noting whether or not the receiver used the same terms for the same landmarks. From Stephen Goldinger, Jennifer borrowed a AXB task used in imitation studies. She found that participants were imitating each other and that this imitation lasted for a while, that is, she found delayed imitation. Goldinger found imitation in non-social settings, and this imitation disappeared after two or three seconds after the opportunity to imitate had gone by.]*

The other study I have done in this domain was with Kevin Shockley at Connecticut. We also used a cooperative task.

*[Professor Fowler describes a task in which two partners must discover the differences between two pictures. The trick is that each participant can only look at one picture; therefore, they must communicate and cooperate in order to discover these differences.]*

*During the experiment, the postural sway of each participant was measured by use of special sensors on the hips and ankles. These were used to measure the changing positions of the body through time. In a 2x2 design, the participants were either facing each other or were positioned back-to-back. They either worked together or with a confederate, that is, they were either cooperating or not. Using cross-recurrent quantification analysis, and applying the embedding theorem, it was shown that postural swaying of the participants corresponded more when they were cooperating than when they were not.]*

In public language use, utterances are used in the context of cooperation. In that context, we imitate one another dispositionally. I think that through imitation we foster inter-personal coordination and cooperation. The two findings are that communicators entrain speech and posture when interacting. Finally, we need a theory of public language use if we are to understand the important role of language in human life. A theory of public language use will constrain a theory about language form. It has been shown how linguistic utterances are interjected in joint activity and how they foster cooperation. Thank you.

#### APPLAUSE

**Robert E. Remez:** Of course we have time for questions.

#### Questions from Audience

**Robert M. Krauss:** Obviously I like this kind of research, but I do have a number of questions. One is about the cooperative, coordinated role of language, which you emphasized. In the literature there are also demonstrations of divergence rather than convergence, and they are found specifically in situations when the relation of the interacting is not entirely cooperative, when it is antagonistic. So what does that tell us? On the one hand you have Herb Clark and others stating that language use, by its nature, is cooperative, and then, on the other hand, you have people saying that language is cooperative but that people can use it to not cooperate. So what is the message that this is carrying?

**Prof. Fowler:** Well, I think that it is both true that language use is by its nature cooperative and that people don't always use it entirely cooperatively. I think that both of those things are true. Even when we're interacting in a competitive or hostile way, we're still speaking the same language, so we're cooperating to that extent. I think that, fundamentally, language use in the public domain is cooperative. But you're right, there are findings of divergence. In our research, it would've been nice to have a competitive scenario in our experiments, but, perhaps, as was found in the Ronald Reagan experiment, no difference would be found.

**Marco Jacquemet:** I would like to continue along those lines. I think that there is bias toward cooperation. I wonder if you ever thought about trying to figure out where the process of coordination happens, and I wonder if it would be better to talk about semiotic understanding, which is that people have to have a shared understanding for language to work. But then, I don't think that at the level of interaction we can assume that people are cooperating.

**Prof. Fowler:** We haven't yet looked at that. The prototypical language activity is cooperative.

**Nina Wacholder:** Would you go back to the notion that perception is attempting trying to go back and figure out what was in the speaker's and head and then, the notion of gesture. Could you expound a little bit on how gesture sheds light on what perception might or might not be like?

**Prof. Fowler:** The thing that the notion of gesture is doing theoretically is that it is defining language forms that are public. Unless we are coerced into thinking about language in the prototypical way, it is odd of thinking of language about having form but not having public appearances. So the concept of gesture is putting the language forms out in public: they are the things that immediately structure the acoustic signal and therefore they are what the signal provides information about. It is a parity-fostering characteristic that facilitates the efficacious use of spoken language.

**Wacholder:** So then perception is going from gesture to the speaker's intention.

**Prof. Fowler:** Perceiving the gesture is the way that it becomes possible for us to determine the speaker's intention. I'm making the claim that language forms are in the vocal tract. It is true that our knowledge of language forms are in our head, but the gestures themselves are public actions.

**Herbert Terrace:** There are many other forms, apart from linguistic ones, that could be said to be in the head. I mean, doesn't this apply to the motor processes involved in things like riding a bicycle?

**Prof. Fowler:** Yes, that is true. I am focusing on linguistic gestures. I think that Chomsky is right: you can't have a theory of public language use, but you can have a theory of joint activity.

**Terrace:** You may know the theory of Merlin Donald, about the evolution of language. He makes the point that language evolved out of joint activity.

**Robert E. Remez:** How is the knowledge needed to detect and imitate a gesture, via an acoustic signal, different from the knowledge that a system would have to have in order to just perceive the acoustic signal?

**Prof. Fowler:** I am not sure. We know that a bilingual speaker, in speaking one language, is influenced, for some times, by the phonemes of the other language.

**Remez:** So that's the type of gestural imitation. Suppose that I wanted to imitate someone. Imitating someone would require knowing quite a bit about that individual. That might be a different kind of imitation. I'm thinking of memory studies. Subjects were asked to imagine a word spoken in either their own voice or someone else's.

**Krauss:** I guess I have a problem. You defined gesture in a conventional way, but speakers imitate more things. For example, they imitate register. Now register is gestural only in the smallest part. So there is a set of rules and conventions that speakers are observing—there are certain words you don't say in the presence of the dean. These are things that have nothing to do with these articulatory gestures. And, of course, the relationship between the gesture and some lower meaning is not straightforward. There are many gestural ways

to realize the same meaningful content. I guess, apart from the part that you are from Haskins, I don't fully understand the reason that you place all this burden on the articulatory gesture concept rather than on something that would allow you to talk about these various levels.

**Prof. Fowler:** I do think that the level I focused on is crucial—they are the language forms, they are the things publicly communicated to other people. But you're right: the full reason why we use language is more than just my work. I am focusing on phonological perception and the language forms.

**Prof. Remez:** Let us thank Professor Fowler and adjourn.

---

**Place:** Kellogg Center, Room 1512  
School of International and Public Affairs  
420 West 118th Street  
Time: 4:00 PM

**Chair:** Prof. Robert E. Remez, Barnard College, Columbia University.

**Attendees:** Melanie Degeratu, Pablo Duboue, Aili Flint, Elizabeth Henly, When-Chia Hu, Boris Gasparov, Marco Jacquemet, Mobina J. Khan, Johnne Kleifgen, Robert M. Krauss, Jason Kruk, Janet Metcalfe, Ezequiel Morsella, Smaranda Mureson, Katherine Nelson, Robert E. Remez, John Saxman, Sarah Schmeidler, Ann Senghas, Valery Shafiro, James Shaw, Anja Soldan, Lisa Son, Andrew Sunshine, Nina Wacholder.

**Rapporteur:** Ezequiel Morsella.





26 OCTOBER 2000

---

## The perceptual organization of speech

Robert E. Remez  
*Department of Psychology*  
*Barnard College*

A fundamental perceptual function evident in spoken communication is the analysis of sensory samples of speech. However, perceptual analysis of the phonetic properties in stimulation cannot proceed as if sensory activity stems from speech sources alone. We speak and listen to each other amid multiple sources of sound. Indeed, the vocal structures are a source of respiratory and ingestive sound as well as speech. In consequence, the perception of speech entails two functions: 1) an organizational function that identifies a sensory pattern issuing from a spoken source; and, 2) an analytical function that identifies the linguistic form conveyed in a sensory pattern. Our studies have evaluated the present dominant account of perceptual organization, which is based on principles deriving from Gestalt accounts of the composition of forms from the elements of sensation. Through acoustic analyses and empirical studies of the perceptual resolution of speech signals, our research motivates an alternative conceptualization. A review of the evidence shows that linguistic perception rests on a perceiver's susceptibility to speechlike auditory variation within broad tolerance for the specific sensory components of the stream; and, depends neither on the apprehension of speechlike auditory qualities nor on the binding of similar, typical or familiar auditory sensory elements into a perceptual stream. The key perceptual phenomena will be illustrated in listening examples.

This whole thing started with an intriguing phenomenon. We wanted to understand how well someone could extract linguistic variables from an auditory signal that matched speech in the most coarse grain way. Here are the items that intrigued us. See if you can make out this sentence.

*[Professor Remez plays an audio sample of a mechanical-like voice saying, 'Where were you a year ago.' The audience can easily comprehend the synthetic sentence.]*

Here are the three time-varying sinusoids that in aggregate compose the sentence you just heard.

*[Professor Remez plays what sounds like a series of muffled, incomprehensible, low pitched sounds. Then he plays the middle frequency component of the sound sample, which sounds like someone whistling, and the high frequency component, which sounds like a high-pitched bird call.]*

Interestingly, when you put all these three incomprehensible sounds together, the message is clear. Why should that be? This is a question about the first 100 ms of speech perception, and much of what sets the boundary

conditions on this problem is the urgent requirement to make something out of an impinging sensory assortment. Those of us who have worked on artifacts that can sample an acoustic array—and hold it as long as the plug stays in the wall—don't face the same requirement; therefore, when we work in that artificial domain, our speculation is much freer than in the domain human psychology. We need a story about what the listener does very briefly—to take impinging sensory elements and to identify them as speech.

"The most important information we can possess is the knowledge that the message we are reading is not gibberish." In this epigram Norbert Wiener clearly presents the problem of decoding. In perception this has a clear correlate. When a sensory domain is coherent, whether that coherence is detected or imposed by a perceiver, then perceptual analysis can ensue. There is no analysis without a well-formed sensory domain to subject that analysis to. In speech perception, this can take place only after the perceiver has detected the sensory manifestation as a contour that stands out from the sensory background. How do you do this? When Ira Hirsh wrote his magisterial chapter on auditory perception for Stevens' Handbook (1988), he put the problem this way: "If indeed the perceiver calls into place a system that is specifically designed to process speech and calls for another system to process non-speech sounds and foreign languages, then in some early level be sufficient auditory analysis to permit the listener to know when to turn on the speech perception system. The crucial experiments to decide this point have not been done." Of course he was being too modest. Fifteen years earlier he and Julesz wrote a chapter in which they did outline a program of research to do the experiments. There had not been any test for speech, but we, as a field, knew quite a bit about how perceptual organization occurred within the auditory domain. In fact, here is a generalized account of the problem of perceptual organization with the specific instance of the talker. There are four domains: 1) a domain of objects in the world, 2) a domain of the perceptual medium that conveys correlates of the object and the event to the perceiver, 3) a sensory domain that elaborates the media as projected sensory commodities, and 4) a perceptual domain, which is structured in an isomorphism with the object domain.

Let's take vision for example. The talker reflects light, which is a perceptual medium whose properties are correlated with the physical and functional properties of the speaker, such that the speaker can be visually differentiated from other speakers. Upon hitting the retina, the light causes a cascade of sensory neural activity. This cascade, in the sensory psychology laboratory, is described as a time-varying impression of hue saturation and brightness which coalesces into an impression of murky blobs suspended in shallow depths. Of course, if perception were to stop with the automatic processes, the perceiver would never say, "Oh, that's a talker and it is President Eisenhower describing his golf game." The perceptual domain is required to take the murky blobs suspended in shallow depth and to project them into an impression of a talker speaking somewhere in the environment.

In the auditory domain, the talker is a source of acoustic patterns transduced in the auditory system that create an impression of a time-varying localized stream of pitch, loudness, and timbre changes. Those are projected

into an impression of the message and also of the person talking. Mechanical vibration and olfaction are also taken into account when attributes of speakers and messages are perceived. (To my knowledge, this is the first time that olfactory experience has been related to the perception of speech.) The formulation of this point of view we owe to Brunswik, whose probabilistic functionalism warranted a program to identify the correlation of sensory elements and worldly attributes. Translating it to our problem, it means that, within the object domain, the talker and the spoken message can be broken down into attributes of talking objects and linguistic objects, each of which has its own sensory correlates—auditory and visual. Of course because there are auditory attributes of visible objects, and visual attributes of audible objects, once these are analyzed, there is the problem of binding the right ones with one another. The combination of these attributes eventually produces a visual and auditory impression of the talker and the message that was spoken.

At the cocktail party, in order to identify the talker as a specific contour amid this auditory world of concurrent sound—of champagne, of cutlery, of other talkers—one has to be able to tell that the specific set of visual and auditory features pertain to a single talker. The case of cocktails for two is more intimate—when there is only a single talker. But, you still have to identify the attributes of the single talker as a coherent set. A related question, one that sets fire to the barn, pertains to the detection of coherence across sensory modalities. This ultimately will undermine any attempt to create a general theory of perceptual organization that applies only within modalities.

So how do accounts of speech perception describe organization? This is an interesting state of affairs. In a catalog of speech perception theories compiled by Dennis Klatt, that these theories express a benign indifference to the problem. Let's consider three types.

In Klatt's portrait of *Analysis by Synthesis*, an approach adopted by engineers, the technique relies on a knowledge base. Theories of this genre assume that there is only one sound in the world—that of the idealized speaker. In these theories, there is no need to find the speaker, for all sounds in the environment are thought to emerge from this idealized speaker. The *Motor Theory* starts with a peripheral speech signal—probably a /ba/, /da/, or /ga/, and, as you can see, through recourse to various motor commands and other forms of knowledge, it comes up with the correct perceptual object, despite all the noise in the world. This is a theory that assumes that there are no errors. The *TRACE* model of McClelland & Elman is a neural net model, though, of course, there aren't any neurons in it at all. It is basically a discriminant analysis for correlating acoustic with linguistic properties. Mainly, these models are indifferent to the problem of organization, presupposing the function without incorporating it.

Our contemporary default account of perceptual organization derives from the Gestalt Max Wertheimer, who attempted to identify the properties that arose spontaneously from a stimulated nervous system. His idea was that the sensory array is not transmitted piecemeal to central sensory areas, but is organized. His stand opposed the Structuralists, whom we all read about in our history of psychology course. Wertheimer said that sensory elements are grouped together when they are 1) proximal, 2) similar, 3) continuous,

4) symmetrical, 6) are separated by small gaps (grouped by the principle of closure), or 7) change in the same way. There is also a principle of habit.

*[Professor Remez shows examples demonstrating each of the Grouping Principles.]*

How can these account for speech perception?

Julesz and Hirsh (1972) actually laid out a series of things that might be true of a listener if Wertheimer's principles apply to the auditory domain. In fact, Wertheimer was right. As was demonstrated in a series of experiments, most of which were conducted by Bregman and colleagues, there is grouping by frequency proximity; grouping by similarity in frequency change; by similarity of fundamental frequency; by common modulation; by similarity in spectrum; by closure of interruptions; by common onset and offset; by frequency continuity; and there is grouping by melodic and metric sets. Perceptual effects of habit are uncertain, because it is hard to know how habit could apply to a function that occurs 100 ms after a waveform hits the ear. How could learning apply in that short span? Julesz and Hirsh and others were looking for something automatic that would perform a primitive analysis. There is plenty of time for the intrusion of knowledge once we get to plausible guessing.

This is the way the theory looks based on Wertheimer's approach, and it was described in Bregman's *Auditory Scene Analysis*. A waveform strikes the ear and creates a pattern of auditory sensations. Grouping of the auditory elements occurs according to the principles in the Gestalt set. These patterns are then elaborated according to schematic knowledge. A *schema* is a description in abstract form of the typical or familiar properties of objects of the world. Contours are identified by reference to schema, some for speech and language and others for different classes of sound sources. The perceptual impressions of objects and events are reported to awareness. There is substantial evidence that this description is adequate for some auditory analysis. Warren, Obusek, Farmer, & Warren (1969) combined different auditory stimuli and found that subjects grouped them together according to Gestalt principles.

*[In audio examples, Professor Remez demonstrates that the perception of the order in a sequence of sounds is lost when the presentation is reduplicated rapidly. Each acoustic component, whether noise, tone or buzz, is as long as a syllable, 200 ms.]*

If one considers the acoustic properties of speech, one realizes that the Gestalt grouping principles can not sensibly apply.

*[Professor Remez shows a speech spectrograph and explains its parameters.]*

On the spectrogram, energy changes asynchronously (e.g., formant frequencies); this asynchrony should be sufficient to split them into separate streams. In addition, there are spectral differences that one sees, differences in the mix of periodic and aperiodic constituents. Each of these sets of dissimilar elements should stand out as a separate stream. Lackner and Goldstein put speech sounds in a reduplicative test and found that when you play the speech sounds /bi/ /a/ /gi/ /u/ in rapid succession, subjects have a difficult time determining the order of the sequence. This shows that Gestalt-based fracturing can indeed occur for speech signals, as it does for other kinds of auditory stimuli. However, they concluded by noting that this kind of

organization does not occur with normal everyday speech. We do not form perceptual streams of speech signals by sorting the elements according to likeness (e.g., by consonants and vowels).

Dorman, Cutting, and Raphael (1975) thought that they had a line on this problem. They first established that, if formant transitions are placed between the vowel sounds, the sound sequence was easy to perceive. In their view, continuity can hold speech together. Of course that would be useful only in a language which consists solely of vowels; such a language is unknown to me. So the problem is that, if you have continuity, you see coherence. Unfortunately, in the spectrograph of an ordinary sentence, one doesn't see the continuity.

*[Professor Remez plays two examples, one in which the vowels are steady-state and another manifesting interpolated formant transitions.]*

The problem for us is to test the premise of the auditory model of perceptual organization: that perceptual organization depends upon physical continuity and on specific auditory elements. A path to an alternative view begins with a question: Does the perceptual organization of speech depend on sensitivity to coherent modulation independent of the auditory elements. That is, is the grain of analysis relevant to organization one step removed from the crude auditory elements that were transduced from the periphery?

Here are our empirical goals. We need a test that distinguishes the perceptual effects of elements from those of patterns, and to see whether a listener, given the choice to go with elements or patterns, will express sensitivity to pattern. We also need a test that precludes mechanical and physiological spread of effect at the periphery potentially inducing coherence. And, we need a test of the independence of auditory organization. To eliminate any suspense about the outcome of these tests, we found evidence of organization that cannot be subsumed under the Gestalt rubric. I will refer to that as "phonetic" in order to be conservative. And, we found that these two types of perceptual organization, those that apply the Gestalt principles and those that apply to speech, are independent.

Consider the spectral pattern of a sentence "The steady drip is worse than the drenching rain," and next to it is a pattern of sinewave that replicates the main features of the natural spectrum. As you can see, the tones onset and offset asynchronously and change frequency at different rates. They onset and offset at different points, and there are elements that come-and-go discontinuously. This pattern, if treated to a Gestalt analysis, should be broken into its separate components, but the intact pattern cannot be resampled for expansion by schemas because the sensory form will have decayed long before the schemas can be applied. Because the auditory system purchases its exquisite temporal resolution by massively damping sensory signals, they don't linger very long. The fact that a listener could transcribe a sinewave sentence at all satisfies a primary criterion of our empirical goal: to show sensitivity to patterns independent of their elements.

The remaining basis for grouping sinewave components is the mere similarity in spatial location, but Gestalt grouping could nonetheless occur on this premise. Our first test addresses this issue. In the test we performed, we

wanted to see whether similarity in location is responsible, for the perceptual organization of sinewave sentence structures.

There were four conditions in this test. In the first, a subject heard sinewave sentences consisting of three time-varying tones presented to both ears. The subjects experienced this as a there being a single sound source located at the vertex of the cranium, localized internally. We expected performance to be quite good. In a second condition, the tones were presented dichotically, and the subject heard one of the tone components in one ear and the rest of the tone components in the other ear. If the subject cannot organize these components, then performance should be no better than performance on the input at each ear alone. A nonadditive effect indicates that a subject can organize these sentence components despite the lack of similarity in location. Two controls were used to estimate performance for each individual ear in the dichotic case. The results show that, though performance is poorer in the dichotic case than in the binaural case, it still exceeds the performance predicted on an assumption that the subject perceives the message from the sinewave signal of one ear and adds it to that of the other ear. This is evidence that perceptual organization across ears is based on the coherence in the time-varying components. The question is, What is the subject listening for? In our experiment, it cannot be similarity of any kind.

In another test, we made the dichotic task competitive. In this task, we inserted a supernumerary tone that is a time-inverted version of the tone analog of the second formant. Statistically, the time-inverted form of a tone has the same properties of a proper component of sinewave sentence structure, but because it is time-inverted, it cannot evoke phonetic impressions in the way the normal sinewave replica does. In this task, the perceptual system must reject the time-inverted component, even though its source in space is the same as its temporally veridical version. In this scenario with the three tones, the time-inverted one should be rejected, and the temporally veridical components should be organized, even though they do not come from the same source. It is a difficult test for listeners to do. But, the results were clear. Performance was poorest when the competing tone changed in a speech-like manner; performance was best when there was no competing tone, or when the competing tone was steady state in frequency, or varied in an unspeechlike manner. These findings indicate that the listener is fooled by the speech-like variation in an extraneous tone.

Let us say that there are these two modes of perceptual organization—1) an element-based approach, described well by Wertheimer's principles and 2) another approach that must be sensitive to coherent organization. What is the relationship between these two modes of organization? Both must take place within the first 100 ms after the waveform strikes the ear. Are they alternative, or can the listener sustain both organizations of a single acoustic pattern concurrently? In an article that is about to appear in *Psychological Science*, we posed this question using sinewave words. We asked whether a subject can simultaneously verify the auditory form of a tone that follows the frequency variation of a second tone, and identify the word that the second tone composes in conjunction with the other tones. As you can see, word

verification was very good. There were two days in which subjects said on each trial whether the sinewave word matched a printed word.

Now, remember that if listeners don't know that they are listening to word forms, they will think that they are listening to arbitrary sounds. After two days, we let them in on the secret—that the sinewave patterns composed words. Amazingly, auditory form sensitivity remains acute even when you ask subjects about the word. This experiment shows that we can maintain two organizations of the same signal concurrently, one auditory and one phonetic. It's our suggestion that organization by phonetically based pattern sensitivity of speech occurs during ordinary listening, to detect whether or not the incident elements form a coherent pattern issuing from a vocal source: It is because you cannot type the elements, that you need to type the chain.

Our conclusions are that phonetic perceptual organization appears to be keyed to speech-like acoustic variation independent of short-term spectra. It occurs rapidly and automatically, and it seems to be sensitive to abstract correspondences despite the lack of familiarity with the acoustic elements. Is there any independent corroboration?

Experiments by Bob Shannon are one corroborating source. Using a set of filters to measure the energy within specific frequency bands of a speech signal, these estimates were then used to control the power of a matching set of bandpass filters excited by a noise source. Although the spectrum remained stationary in frequency, the rise and fall of energy was correlated with the natural speech used as a model. When four noise bands change in intensity asynchronously, the sentence pattern was discerned by listeners. Because the noise bands are not changing in frequency, one cannot think of them as speech-like, although they are speech-like in course grain. In the narrow grain at which Gestalt principles apply, though, they are unable to rationalize the grouping of the noise bands into a single speech stream.

*[Professor Remez examples of noise-excited vocoder signals that replicate the temporal properties of the rise and fall of energy within different frequency bands. In one, the spectrum from 0-5 kHz was varied together; in a second, the variation was divided into 4 bands, each approximately 1 kHz wide. Only the latter was intelligible.]*

Another researcher, Bertrand Delgutte, also creates a type of speech display described as a vocoder, in which there are excited bands that vary in gain over the course presentation. Now, Shannon used a flat spectrum of noise as a source of excitation. We could imagine devising an algorithm to calculate course grain representations of the speech signal on the presumption that the source of excitation is a steady pattern of noise. In contrast, Delgutte used a brief recording of a jazz ensemble to excite the filters. If you analyze what you hear, you can hear a saxophone, several drums, cymbals, a bass fiddle, and a polyphonic keyboard (it used to be called a piano) playing *Take Five*. It is a very complicated scene. When a vocoder is excited with this...

*[Professor Remez presented examples of the natural sentence, the sample of Take 5 used to excite the vocoder, and the resulting chimerical signal.]*

The listener had to follow a nonstationary signal in an adaptive, self-organizing manner, detecting the differences between what the signal should

have done if it consisted of instruments alone, and what it did because it was passed through an envelope whose shape was set by a speech signal. This is powerful corroboration that the perceptual organization of speech is independent of auditory organization, that it can be sustained concurrently, and that it is an automatic process that applied without much effort in circumstances that are neither typical nor likely. And, this is the problem of perceptual organization.

## APPLAUSE

### Questions from Audience

**Robert M. Krauss:** In reference to that Norbert Wiener quote and the idea of habit, we have to realize that the two are not unrelated: one simple observation is that one does not have to know a lot to hear a lot. Basically, I'm not sure that I could easily identify something—for example, the song *Take Five* that you played—without having been exposed to it before. Another observation is that, from various studies, it has been shown that infants respond to speech as if they have some sort of propensity for it. Shouldn't your model incorporate some sort of bootstrapping mechanism to account for this type of necessary knowledge, that is, the kind gained from experience and the kind that is inborn?

**Prof. Remez:** One has to segregate aspects of organization that stem from habit from those that stem from innate mechanisms. In reference to the latter, the following questions must be asked. First, when did the first hominid speak? Second, when does a baby first recognize a speech signal? Third, when, in the nervous system, does a speech signal hit the ear? The evolutionary system has to obtain coherence under pressure, so you have to rely on up front processing. Chomsky claims that we have this innate ability to perceive language, that we have this of language, but not for anything else. An experiment by Eimas and Miller showed that infants can fuse dichotically dispersed components, so their must be some rudimentary mechanism in place before learning can take place.

**William Benzon:** It has also been shown that infants display synchrony with the intonational patterns of adult voices.

**John Saxman:** It has also been shown that the auditory system develops very early, in utero. Here is a naïve question: Wouldn't the ability to characterize patterns be fundamental to perception?

**Prof. Remez:** Yes, and the elements must first be decoded as a function of some type of perceptual organization. Before neuroanatomical studies, the debate in the nineteenth century was between the Gestaltists and the Structuralists. The Gestaltists believed that a perceptual system had to be sensitive to a pattern, whereas the Structuralists believed that sensory elements were the units of perception.

**Daniel Ellis:** Are you saying that in the first 100 ms there isn't enough time for a guess, for a knowledge-based process?

**Prof. Remez:** Auditory grouping cannot know what you know, because it operates so fast that it cannot refer to a knowledge base. There are several

points against a knowledge-based description of human perception: 1) Infants don't wield such knowledge, yet perceptual organization occurs despite their immaturity; 2) by the time a schematic process realizes that the primitive parser has committed an error, the auditory trace has faded, preventing reorganization; 3) there is no evidence that a perceiver can learn to revise a primitive perceptual grouping mechanism; 4) schematic knowledge represents the likely or characteristic properties of familiar sounds, which falsely predicts a failure of organization with sounds that are unfamiliar, such as sinewave speech, noise band vocoded speech, or chimerical speech.

**Ezequiel Morsella:** How do you separate the perceptual object from the information requirements of the task at hand? In other words, how do you know whether what you have is a bistable percept, or whether the percept suits itself to satisfy the task at hand, at that instance in which the task takes place? For example, when speaking about a chair, a subject may refer to the structural or functional features of a chair, which could be quite different in description. Does that mean that he has a bistable percept?

**Prof. Remez:** That is an excellent question, but a hard one to answer. In short, our technical assessment of the percept comes through the actions of a subject, and so this is always a profound problem of proof. But, one reason that the intelligible sinewave and chimerical items are so provocative is that they are perceptually bistable, or multistable in a way that an ordinary chair, despite its potential multistability, is not. The multiple potential uses of a chair do not hinge on alternative, incompatible physical descriptions of the object. In the same sense that the Rubin vase cannot both be a vessel and two profiles, the bistable acoustic patterns evoke incompatible impressions. However, in contrast to the visual cases, which are successive and reversing, the auditory cases are simultaneous.

---

**Place:** Kellogg Center, Room 1512  
School of International and Public Affairs  
420 West 118th Street  
Time: 4:00 PM

**Chair:** Prof. Robert E. Remez, Barnard College, Columbia University.

**Attendees:** Peter Balsam, Janelle Barnes, William Benzon, Angels Colome, Paul Currie, Dan Ellis, Jennifer Fellows, Boris Gasparov, JoAnne Kleifgen, Robert M. Krauss, Lance Kriegsfeld, Marlene Lipson, Joseph LeSauter, Michele Miozzo, Ezequiel Morsella, Rebecca Piorkowski, Jennie E. Pyers, Abe Rosman, John Saxman, Sasa Schneider, Harriet Tabre, Herbert Terrace, Harrison White.

**Rapporteur:** Ezequiel Morsella





30 NOVEMBER 2000

---

## The cognitive consequences of linguistic change

William Labov  
*Department of Linguistics*  
*University of Pennsylvania*

No one doubts that human language is designed on the whole to deliver information; when interlocutors misunderstand one another, and wrong information is transferred, some explanation is called for. The traditional view of language change, and sound change in particular, is that it seriously interferes with this primary function. If this is so, it becomes urgent to re-address the long-standing problem of the causes of linguistic change. The recently completed *Atlas of North American English* has confirmed early indications that sound change is continuing at a rapid rate in all the major cities of this continent, and that regional dialects are becoming increasingly differentiated from each other. Many of these sound changes are mergers, which reduce the information-carrying capacity of language. Others are chain shifts which by definition preserve distinctions and avoid increases in homonymy. The project on Cross-Dialectal Communication addressed the question as to how ongoing chain shifts affect the efficiency of communication across regional boundaries. Gating experiments were carried out in Philadelphia, Chicago and Birmingham, representing three radically different vowel systems which are changing in opposite directions. Results show that the advanced forms of the chain shifts are frequently miscategorized by listeners from other regions, but they also create misunderstanding within each speech community as well. The reverse situation can be argued to hold in the case of mergers. In her study of the widespread merger of long and short open /o/ in *cot* vs. *caught*, etc., Herold maintains that mergers can be viewed as an increase in information rather than a decrease. The study of a large body of natural misunderstanding confirms Herold's theory, showing that speakers from communities that have undergone merger rarely misunderstand speakers from communities that have maintained the distinction. The consequences of these findings for our general understanding of the functions of language will be considered.

I am going to begin by reading a little bit from the beginning of something that will appear in a forthcoming book on the principles of linguistic change. The first chapter is called "The Darwinian Paradox." The main idea is that language change and biological change are quite parallel in their formal appearance, but the crucial distinction is that adaptation and natural selection does not apply to language. We are all still very much puzzled by that. To put it simply, what good does it do us to be unable to understand the reference?

[Prof. Labov reads aloud an excerpt from the book. A paraphrase follows.]

Each of us has suffered the effects of language change in one way or another. These effects range from petty inconveniences to crushing

disabilities that could consume years of our lives. The ebb and flow of the properties of words pose difficulties. My generation called an ice box an *ice box*, because it used a block of ice before it was electrified, but my kids' generation insists on calling it a *refrigerator*. I am also at my age ridiculed for calling something *suave*. On the other hand, there are forms of the language that people object to, like *ain't I* instead of saying the proper *aren't I* or using *like* as a conjunction. Even the most strict educators cannot prevent these linguistic events. These so-called defective forms exist again and again until they change the fabric of the language.

On the other hand, most of us have suffered by having our school papers downgraded by not following the conventions that preclude ending a sentence with a preposition or forgetting to place an *m* at the end of a *who*, for some reason. These disagreements are minor inconveniences. When we observe such controversies in foreign languages, it is easy enough to see them as tempests in linguistic teapots. But within our own language it is difficult not to get caught up in the spur of the emotions generated by the contrast between newer and older versions of a word meaning the same thing. It is not easy to step far enough to ask the fundamental question: Why does language change? A great amount of our time and effort is devoted to mastering English spelling, for example—words like *bite*, *drought*, *might*, and *draft*. With the alphabet I find myself sometimes fighting an uphill battle, distinctions like *wear* and *where*, *morning* and *mourning*; and *kernel* and *colonel*. A generation ago, people produced and heard a difference between *morning* and *mourning*, and *kernel* and *colonel*. So there are great disadvantages resulting from language change: you cannot learn to read English without investing two or three extra years to master spelling.

Some people say that there is a positive side to language change, that learning several languages renders someone cognitively superior, but this has not really been proven. It is hard to avoid the conclusion that language, as an instrument of communication, would work best if it did not change at all. We all seem to be suffering from a linguistic disease that has no cure. Language, like much of the world around us, seems to be going from bad to worse. The Golden Age principle is usually applied to music, fashion, and architecture. But it applies most to language. Sometimes you hear older people say that they like today's music, dress, or computers, but you never hear someone say, "I love the way young people speak today!" [*The audience laughs.*]

If it is really true that language change interferes with communication, then what are the factors that produce language change? How do we reconcile language change with our fundamental view that language is an instrument of communication which differentiates us from animals—who have various systems for communicating local identity, aggression, and submission, but not information about when they are going to meet someone downtown or what time of day it is.

[*Prof. Labov shows a slide of a map that will appear in a CD ROM that will accompany the Atlas of North American English. The CD ROM map was prepared by collaborators at the University of Auburn.*]

This *Atlas* is a project that I started in 1994–1996. There has never been an atlas of the pronunciation of American English until this one. All European countries have maps for their languages, two or three generations of atlases,

and there are many reasons for this. The reason a map is so important is that there is tremendous language change in the urban centers of this country, so that a study on language pronunciation conducted in the 1940s will have results that are quite different from that of today, because the language changed so much. Most the linguistic changes that we are looking at take place at a very fast rate in New York and Chicago, or Buffalo, and they stand for two or three generations at the most. So we need a rapid method of data collection, and it occurred to me early in the game that telephone is the way to do it. A telephone survey could cover all of the United States in three years, and that is what we did. Our study is not of all the areas of the United States, but of the urban areas, in a technical sense, and this represents about 60% of the speakers in the population. We do this by calling people up who are local, that is, who are born and raised in the region of interest. Now, in order to represent a city like New York City, you need 120 speakers to represent the people of the city; but to represent the continent as a whole, we need from each area only four or six talkers. As you will see, this was successful, and one of the biggest puzzles was to discover why this method was successful.

*[On the projection screen, Prof. Labov shows the table of contents of the Atlas.]*

What I'll be talking about today are *mergers* and *chain shifts*, because these are two different kinds of language change in regard to the fundamental question: does language-change interfere with communication? Mergers represent a collapse of a distinction, so that *pin* and *pen* collapse and become the same. We have lost the capacity to distinguish between *him* and *hem* and *since* and *sense*, among other losses of information. Presumably this will lead to an interference with the communicative capacity of the language. On the other hand, chain shifts represent the opposite. A chain shift, like the great vowel change in English, is a preservation of distinctions—one vowel starts to move, and the others rotate behind it in a game of musical chairs that preserves the number of categories and sounds in the language. Therefore, there should be less communicative interference.

*[Prof. Labov again displays the map stored on the CD ROM. He adds that, once released, the CD will allow people to hear all the dialects from all parts of North America—Canada and the United States.]*

You could see that there is a merger that exists in most of the South and extends beyond the borders of the South. As a principle, mergers will expand at the expense of distinctions. That principle was developed by my friend and associate, Marvin Herzog who has recently succeeded in publishing a language and cultural atlas. Herzog showed that this principle is a consequence of one of the fundamentals of linguistics: the arbitrary character of the sign. Things could be called anything, but the average person doesn't believe that, he believes that there is an actual way to speak about a pig and that's *pig*. *[Laughter from the audience.]* Because the sign is arbitrary, it turns out that, when two categories of sounds come together as one category, it really becomes impossible to separate the categories again. If you take a person who has learned a merger, though the spelling of the words remain different, he or she will find it impossible to learn the previous distinction. Then again, if you learn a distinction you could

easily learn to omit it. Herzog raised the question, What happens at the intersection of the dialects? Logically, it should produce an unacceptable result: All the vowels are pronounced the same way. In fact, that is what happens: the mechanical aspects of the merger triumph over the importance of communicating information.

*[On the map, Prof. Labov points out regions where mergers have taken place, mergers like caught and cot sounding the same.]*

Then there are areas that resist the merger. As you will see, there are three different ways that you can insulate your child from the pernicious effects of a merger and protect him from the collapse of *caught* and *cot*. You could be raised in the area from New York to Baltimore. You are protected by the fact that the *caught* has the high vowel that you have come to associate with the Northern states. The Southern region is insulated by a completely different strategy; it has the vowels sounding the same position for the nucleus, but the *caught* has an additional back upglide. This preserves the distinction and prevents the merger.

You get some idea of the sweeping character of this merger phenomenon. We have some areas of the merger, some areas of resistance against the merger, and some areas of transition, where we find all kinds of intermediate phenomena. Now, what causes them? I said before that the expanse of the mergers is a consequence of the fact that *it is possible to lose a distinction, but impossible to gain one*. We can go into this into much greater detail. I would like to show you an area of Eastern Pennsylvania, where a merger has been expanding. In Eastern Pennsylvania we see the comparison of the 1940 and 1948 isoglosses, of the *cot-caught* merger. Western Pennsylvania had the merger, but it expanded easternmost in 1988. This was learned by telephone survey. However, Slavic immigrants influenced the language change in some regions of Eastern Pennsylvania, so the merger was affected. Most of these Slavic immigrants worked in the coal mines, and in areas like Hazelbury, and Pottsville.

*[By referring to the map projected on the screen, Prof. Labov describes the language characteristic of related areas nearby.]*

In 1977, a group of us went out to Eastern Pennsylvania to a number of towns, and in a recreation area, we interviewed boys and girls. As you see, 17% of the girls said that *cot* and *caught* were the same, in speech and in judgment; 29% of the boys thought they were the same. In 1988, 100% of the girls thought that *cot* and *caught* are the same; the boys thinking this increased similarly in percentage. Again, perception is outrunning production. The merger had a dramatic explosion in a period of eleven years. The speed of this process can be very great. How can the language suffer such terrible deterioration? Ruth Herold, the person in charge of this research, says that, contrary to what we may believe, a merger is not the loss of information, but the gain of information. This is a great paradox. How can she defend this point? Well, she says, you observe that, when a dialect with the distinction is in contact with a dialect without the distinction, the people who make the distinction frequently misunderstand the ones who do not—but not *vice versa*. After a while, the

people who make the distinction (between *cot* and *caught*) realize that they cannot understand the ones who do not, and they come up with a technical reaction known as “the hell with it.” [*The audience laughs.*] They do not necessarily stop producing the distinction; they just do not use it for semantic interpretation, which is the intelligent thing to do.

Does she have any basis for this view? Well, for a long time, in our studies of cross-dialect communication, we have collected natural misunderstandings observed during everyday speech. Observing those misunderstandings gave us a large body of data with which we could test the hypothesis that Herold was putting forward. Most of the misunderstandings are quite strange and funny, but they are not relevant to our discussion because they are basically the result of a deteriorated linguistic environment. We are interested in when people clearly hear what was said and yet still misunderstand—that’s typically because a phonemic segment is misunderstood, e.g., hearing *copy machine* when “coffee machine” was said, or *copy shop* for “coffee shop.” The mechanical character of these misunderstandings show that we never learn, that we never learn the distinction. The deficit is only for those who have the distinction. Therefore, this gives some strong confirmation for Herold’s view that the strength of the merger is an actual gain.

Leaving mergers, looking in the South, we see the Southern Shift. The Southern shift consists of two processes: 1) the sound /i/, as in *sleep*, is lowered to /ei/, as in *hay*, and 2) and /a/ becomes /ai/ as in *eye*. The whole thing is triggered off by the monophthongization of /ai/, as in *eye*. So *seek* becomes *sake*. In the meantime, the short vowels—short *i*, short *e*, and short *a*—are raised to the position that these vowels would formerly occupy, but they are distinct because they have an in-glide. In an experiment we recorded voice samples from speakers of the Birmingham area (in the South), extracted the leading exponents of these sound-changes, and played words samples to people from other areas of the country and asked them to transcribe them. The speakers were college and high-school students from the area.

*[Prof. Labov then played an audiotape of voice samples that were played to speakers from various areas of the United States—Birmingham, Chicago, and Philadelphia.]*

This is one of the methods we employed to see if sound changes interfered with communication. In more northern speech, /ei/ is up and short *e* descends. In Tennessee, we can see how the /ei/ has descended to a lower position overlapping with /ai/. In the meantime, short *i* and short *e* go up. This map gives you an idea of the geographic distribution of this phenomenon

*[Prof. Labov shows a map of the distribution of vowel contrasts in the Northern United States.]*

This is a thematic map showing the percentage of the monophthongization of /ai/ per region. As you can see, in the majority of the cases, it is 100%. A lot of the people in the Greenwoods do have monophthongization, but not before stops. Within the region, there is also a second sound change. In summary, the South represents a coherent view of the progress of sound change, and the as a result of these sound changes it is difficult for people outside of the South to

understand Southerners. Now that might not seem to be important. [*The audience laughs.*]

Here are some of the results of our experiments. We see that, in transcribing, Birmingham speakers are better at transcribing Birmingham voice samples than Chicago speakers are. Only 40% of the people from Chicago could transcribe these words accurately. However, interestingly, only 40% of the college students from Birmingham could not identify their own speech. You could see that, with context, the Southerners are near 100% accuracy; but outside context, they are not. We also see that the sentence context has an effect.

The phonological system is supposed to work this way. Imagine that you are at a party and meet someone who says he is called "Miller." Immediately, you know that his name is not "Diller" or "Tiller." It is Miller. Now, the knee-jerk interpretation is that context always tells what the difference is. But this is not true, because context does not always tell you—even when the context is quite clear, people still cannot perceive the word accurately if it is not produced in the known manner. Of course people from Birmingham do better when the target word is in context than out of context, compared to people from Chicago or Philadelphia. Nevertheless, the people from within the community are suffering from the effects of language change as are the people from outside the community.

To see how this works, we will go on to look at the Northern City Shift, which is a dramatic pattern that started in the 1940's through the 1950's. In 1969, Benjamin Wald and I elaborated it. In Detroit, the /æ/, as in *bat*, has risen to the point that it is higher than /ɛ/, as in *bet*. In the meantime, her short *o* has moved forward to a central position, and short *e* has moved backward; wedges move back. In Wisconsin, we find advanced speakers of the Northern City Shift. Short *a* has moved very high in front. Even phoneticians, who hear these samples as clearly as possible, find it hard to believe that these sound changes occur.

*[Prof. Labov plays many audio samples of the Northern City Shift. One sample shows that the Northerners say black they way we New Yorkers say block.]*

Interestingly, in the South we see that the best transcribers are the high-school students. But then, the college students are worse. Interestingly, however, the high-school students are better at transcribing the individual words, that is, context-free; the college students are better at transcribing from context. This makes sense. There are things you lose while going to college. [*The audience laughs.*] College students are in a formal testing situation, and so the language of their college context influences their judgment. As well, Black students were better than whites at individual words.

The Northern City Shift is the most profound change in English phonology in the last 800 years, because the short vowels of English were stable before this. For some reason, somewhere in the middle of the twentieth-century, speakers in certain areas of the United States began to rotate these vowels in such a form that their speech is incomprehensible to people outside. You may say, "I understand people from Chicago." The question is, Do you really? There is a strong tendency in speakers of English and any other language to

exaggerate the extent to which they really do understand others. One-third of the natural misunderstandings that we have gathered are due to dialect differences.

*[Prof. Labov quotes many examples of misunderstandings across English dialects, misunderstandings cause by the Northern City Shift—e.g., someone hearing pot instead of pet.]*

Now, we need measure of the Northern City Shift. We get that by looking at the mean differences of the second formants of short *o* and short *e*. The interesting thing about the Northern City Shift is that, although the pattern emerged in around the 1950s, the immigrants responsible for the change arrived a century ago, to Syracuse, Green Bay, Rochester, Buffalo, Gary, Chicago, and Flint. These northern areas were inhabited by Germanic and Danish immigrants who were interested in farming. The immigrants that influenced the New England and some southern regions, however, were more religious and ideological. There are two streams in the North: the New England Stream and the Pennsylvania Stream, and these two groups not only have phonetic differences, but ideological ones as well.

What are the causes of sound changes? One answer appeals to the notion that there are universal principles of language change—long vowels will rise, and short vowels will fall. Why are these changes regionalized? They are regionalized because the vowel of one region may be moving in a different direction from another region. What makes them move? Well, they are moving until they encounter another system that is moving in another direction.

As we can see, language change is a natural process that counters a Darwinian view that the properties of language serve communication. However, the phenomena are not so straightforward when we consider things like Herold's claim that a merger is a gain of information.

## APPLAUSE

### Questions

**John Singler:** It seems that advocacy of the death penalty may be related to the ideologies that accompany the Southern linguistic group, but then again, most of these measurements were in cities where murder rate is high. What about rural speech?

**Labov:** Rural speech was not included, but what is interesting is that the larger the city the larger the sound change. But it is opposite in the South. The Southern shift is not advancing as shifts advance in other regions. In the South, the smaller the area, the larger the change. There may be different mechanisms of change in rural and urban regions.

**Robert M. Krauss:** You started speaking about lexical changes, but you said sound change is a structural feature of the language. What drives lexical change?

**Labov:** That has troubled me for a long time, and not much is known about lexical change. I do not think we can say something until we study slang. We know that there are things like *lexical repair*. There are also things like saying *pop*

for *soda* or *coke* for *soda*; and that there are twelve words for language. Also, we know that small dialects disappear. The question is whether there is a parallel between sound and lexical change.

**Benjamin Wald:** What is happening to the Northern England dialects?

**Labov:** I really do not know of any new changes.

**Benjamin Wald:** Why are there two linguistic areas in the North?

**Labov:** It used to be a continuum, but there is a special city that doesn't really fit in with the two regions in a predictable way: I am speaking about Erie. Erie changed allegiance from one group to another. This is because it was a travelling point—you had to stop at Erie to get to Buffalo; it was also a vacation place.

Thank you so much for having me.

[Applause.]

---

**Place:** Kellogg Center, Room 1512  
School of International and Public Affairs  
420 West 118th Street  
Time: 4:00 PM

**Chair:** Prof. Robert E. Remez, Barnard College, Columbia University.

**Attendees:** Elissa Austria, Maryan Baleht-Rofheart, Lila Braine, Cate Crowley, Komlan Essizewa, Rosette Finnerman, Aili Flint, Roger S. Frantz, Stephanie French, Bill Hadican, Elizabeth Henly, Kate Hoffman, Franklin Horowitz, Johnne Kleifgen, Yukiko Koizumi, Paula Korsko, Robert M. Krauss, Santoi Leung, Jeongwan Lim, Oliver Mann, Marissa Montelro, Ezequiel Morsella, Fernando Naiditch, Susan Nakamura, Katherine Nelson, Rebecca Piorkowski, Cristina Rosado, Paula Rubel, John J. Sidtis, John Singler, Joowon Suh, Patricia Sweeting, Diana Van Lancker, Benjamin Wald, Harrison White, Marcie Williams.

**Rapporteur:** Ezequiel Morsella.



25 JANUARY 2001

---

**A deficit in visual location perception:  
implications for spatial representation  
and reading**

Michael McCloskey  
*Department of Cognitive Science  
Johns Hopkins University*

A university student with no history of neurological disorders is profoundly and dramatically impaired in perceiving the location and orientation of visual stimuli. Her deficit is highly selective, and her errors are highly systematic. I discuss the implications of AH's performance for issues concerning the representation of location and orientation in the visual system. In addition I describe a series of experiments that explored AH's reading performance. Reading of text appears relatively intact to casual inspection, but AH is severely impaired in reading isolated words and sequences of unrelated words. I present evidence that her reading impairments result directly from her visual localization deficit, and that her reasonably intact reading of text reflects on-line compensation for the perceptual deficit. Finally, I consider the implications of AH's performance for understanding of developmental reading deficits.

Paying attention to visuospatial information is required for just about everything—navigating through space, picking up objects, reading, and writing. Today I will tell you about a woman with deficits in visuospatial perception. We will see that systematic analysis of the visuospatial deficit gives us a better understanding of the deficit and of normal, intact functioning. The patient will be referred to as AH. AH has normal sensory function, no diseases or injury, and, apart from her deficit, is normal in everyday performance. In her life, she has had trouble with reading, spelling, and math, and, as we will see, this stems from her visuospatial deficit.

When AH is asked to copy a line-drawing, she often draws the image in reverse. For example, when asked to draw the symbol  $<$ , she will sometimes draw  $>$ , a reversed image of the original. She makes errors even with very simple stimuli, such as a curved line. She also has problem in determining the locations and orientations of stimuli. In one task, she was asked to move a computer mouse so that a pointer on the computer screen would touch an X. She was very bad at this task, and her errors were systematic. She has similar problems with orientation—she was wrong in judging the orientation of an arrow, even when there was no time pressure. The question is, How does she reach for things in everyday life?

In one task, she had to reach for a wooden block after closing her eyes. After opening her eyes, she had to name the shape of the block and reach for the object.

*[Shows a schematic depiction of the task and a videotape showing AH's difficulties with the task.]*

AH was wrong about the location of the object on 50% of the trials. Her movements were not random, but usually the mirror reverse of the object. So how does she get along in everyday life? Interestingly, AH thought that her behavior was normal—that nothing was wrong with her. Well, it turns out that there are compensatory strategies that she employs. She continually corrects her movements. These compensatory strategies enable her to get through in everyday life.

The deficit is strictly visual, because she has no problem with determining localization by use of other modalities (e.g., hearing). But with visual stimuli, she always has problems, whether in reaching, grabbing, drawing, or speaking. The type of response does not matter.

We also believe that this is a developmental deficit, not an acquired deficit. We believe it is developmental because there is no evidence of brain damage, and because there is evidence that she had difficulties in her childhood. Her mother saved sketches she made as a child. As you can see *[showing the sketches on an overhead projector]*, the sketch of this Renoir painting is the reverse of the original. Also, her difficulties in math could be explained by the fact that she reverses the numbers. Again, the errors are systematic.

Looking into her deficit a little closer, we see that her performance is affected by some visual variables. One variable is *exposure duration*, which we varied from 17 milliseconds to 1 second. As you can see on this graph *[shows graph plotting accuracy as a function of time]*, error rate increases with exposure duration. Her error rate increases to 50% when the exposure duration is 250 milliseconds. In contrast, at 50 milliseconds, she makes 0% errors.

A second variable that affects performance is *motion*: stationary objects are the most difficult for her. She performs better when the stimulus is moving. A third variable is *flicker*. Flicker improves performance. On a *Visual Retention Test* her performance improved from a 70% error rate to 0% when the stimulus was flickering. A fourth variable is *contrast*, with low contrast being better than high contrast.

The effects of these variables could be explained by appeal to the effects of multiple visual subsystems, each specialized for different stimuli. Here we have two subsystems: 1) a *transient* system which is sensitive to rapid change, and, 2) a *sustained* system which is sensitive to long duration. This is not exactly the same as the distinction between the magnocellular and parvocellular layers in the later geniculate nucleus of the thalamus. Both of these systems are involved in locating and knowing the orientation of stimuli. The sustaining system is systematically full of errors. AH is systematically misperceiving the stimuli—she reports what she perceives. Also, I should mention that the famous WHAT/WHERE and WHAT/HOW distinctions that have been spoken about in vision do not explain her deficit, because there is evidence of overlap between the two systems (WHAT/WHERE, or WHAT/HOW) in her behavior.

Now let us look at her reading performance. Obviously, if she has trouble with object location and orientation, she should have difficulty reading, because reading requires such visuospatial abilities. To read, one needs to know the order of words and letters. AH claimed not to have any problems reading.

*[Shows a videotape of AH reading a short passage.]*

Although she seems normal when reading a passage, she made 25% errors when reading individual words. For example: for “pen” she read “den;” for “lamp” she read “lamb;” for “duck” she read “buck;” for “snail” she read “nails;” and for “nose” she read “hose.” Interestingly, she has problems with sequences of words that do not make sentences.

*[Plays video of her reading word sequences in the wrong order.]*

**Comment from audience: Dr. Jennifer Mangels.** It seems she gets worse toward the end of the sequence of words.

**McCloskey:** That was just an accident. Actually, when reading a text, she is worse at the beginning.

Explaining her reading impairments as resulting from a visual-spatial deficit tells a nice story, but is it right or wrong?

In a series of experiments, we tested whether her reading of single words was affected by the variables of duration and flicker. Again we see that when the stimulus, in this case a word, is presented for a long time, the error rate increases. As well, when the stimulus word is presented with a flicker the error rate decreases from 23% to 1%.

In reading a sequence of unrelated words (2 word sequences), we see that performance again is inversely related to exposure time. When the sequence was presented for 1500 milliseconds, the error rate was 26%. When the presentation was brief, the error rate was only 2%. Flicker, again, affected performance, with the error decreasing from 84% to 3% when the sequence was presented with a flicker.

**Comment from audience: Dr. Jennifer Mangels.** Are there more errors in the middle of the sequence? Does she ever mirror-reverse the items?

**McCloskey:** She switches items that are close to each other. The important thing is where attention is guided.

**Comment from audience: Dr. Sam Glucksberg.** Are all the errors meaningful?

**McCloskey:** Actually, knowledge-based constraints are what allow her to get by. There are such constraints in orthography and in discourse. For example, if you read “the horse’s hoo,” it is easy to guess that the statement should read “the horses hoof.” This is the type of strategy that AH employs. This type of knowledge compensates for the poor visual information. So we looked at what happens to AH’s performance when knowledge-based constraints are missing. For example, in the sentence, “The Smiths moved from New York to Pennsylvania,” there are no constraints regard the two cities. In this case, if AH

switched New York and Pennsylvania, there would be no way for her to detect that there was an error—again, because there are no constraints.

When reading an unconventionally “meaningless” paragraph AH performs worse.

*[Prof. McCloskey reads aloud a paragraph that is quite nonsensical. Prof. McCloskey shows a videotape of AH reading the paragraph.]*

We monitored her eye movements. As you can see, her performance is quite poor—there are even transpositions across lines. Interestingly, AH was better at correcting errors than controls. She corrected up to 85%. If the paragraph flickers, she makes no errors.

**Comment from audience: Prof. Sam Glucksberg.** How did she learn to read?

**McCloskey:** She had no difficulty learning how to read. She is also a very bright girl. She is aware that she has errors, and sometimes she can get quite frustrated. But subjectively, she is not aware. When reading Shakespeare aloud in high school, her teacher told her, “Please don’t correct Shakespeare.” *[Audience laughs.]*

The implication is that from AH we can learn about normal cognition. By applying cognitive methodology, we can learn about her deficit and about normal performance. For example, we see that developmental reading disabilities can have many causes, not just phonological or congenital ones. The more general lesson is that the same deficit can have different etiologies. Developmental dyslexia may be more than just one entity.

## APPLAUSE

### Questions

**Prof. Krauss:** How would she perform on a “same or different” task?

**McCloskey:** Interestingly, stimuli which are the same are never seen as different, but stimuli which are different are sometimes seen as the same.

**Prof. Senghas:** How does she perform on an eye-tracking task?

**McCloskey:** She performed a basic saccade test and had some problems. Also, she blinks more than normals, which is interesting. She also has spelling errors, with a strong tendency to duplicate letters.

**Prof. Matin:** Are the errors the same for each eye?

**McCloskey:** Error rates are the same for each eye—the deficit is at a higher level.

**Prof. Matin:** It reminds me of the prism adaptation studies.

**McCloskey:** That’s a good point, but it appears that in her condition, perhaps because the problem is not at the periphery, there is no adaptation.

Her error rates are definitely affected by attention, which is an interesting variable. Her focus of attention affects the coordinate system.

**Prof. Metcalfe:** Could the side she focuses attention on predict error type?

**McCloskey:** These effects are not taking place at an early point of visual perception, so you cannot predict errors by where she foveates.

**Mr. Morsella:** If she reads the word “dog” as “bog,” do you still get semantic activation from “dog,” that is, does “dog” enter the semantic system? I imagine that this could be tested in a typical priming paradigm.

**McCloskey:** I think that “dog” doesn’t enter the semantic system because the error occurs before then.

**Henry Kong:** Are there any other cases? I imagine that a sample of such subjects would allow you to localize the lesion.

**McCloskey:** Unfortunately, she is the only case I know, though I heard of a few others.

### APPLAUSE

---

**Place:** Faculty House  
Columbia University  
400 West 117th Street  
**Time:** 4:00 PM

**Chair:** Prof. Robert E. Remez, Barnard College, Columbia University.

**Attendees:** Simon Fischer-Baum, Aili Flint, Elizabeth Freidin, Elena Hontoria, Boris Gasparov, Sam Glucksberg, Brian Jacobs, Marco Jacquemet, Mobina J. Khan, Johnne Kleifgen, Henry Kong, Lauren Kornrick, Robert M. Krauss, Claire LePichon, Wenxun Li, Taosheng Liu, Jennifer Mangels, Douglas Meehan, Janet Metcalfe, Ezequiel Morsella, Rebecca Piorkowski, Robert E. Remez, Sarah Shuwairi, Ann Senghas, Anja Soldan, Andrew Teich, Bhavana Vishnubhotta.

**Rapporteur:** Ezequiel Morsella.





22 FEBRUARY 2001

---

## Infants' segmentation and recognition of words

Peter W. Jusczyk  
*Departments of Psychology and Cognitive Science*  
*Johns Hopkins University*

Many recent studies indicate that infants begin to segment words from fluent speech during the second half of their first year. Low-level speech cues are used as a basis for locating the onsets of words in fluent speech. Information about which syllables of words are typically stressed, which sequences of phonetic segments typically occur together, etc., seems to help infants identify possible words in their language. Moreover, infants appear to encode information about such sound patterns in long term memory, suggesting that they begin to build a lexicon during the latter half of the first year. One important difference between infants and adults is the extent to which each is able to draw on knowledge of words to help in word segmentation. Infants know fewer words, and rely on low-level speech cues more than adults do. Because one goal of studying the early development of word recognition abilities is to understand how such abilities evolve into those found in adults, it becomes important to ask how and when infants begin to use information about the words they know to discover new words. I will present the results of several recent studies with infants from 12- to 24-months that have begun to address this question, using a variety of testing methods.

What I am going to tell you about today is the beginnings of word segmentation and the beginnings of word recognition in children. I will first try to convince you that infants have to segment words in the first place. There are many reasons. If you exclude vocatives, fillers, and greetings, only about 7% of utterances addressed to infants between six and nine months of age consist of isolated words. This is a heroic study by van de Weijer, who followed and recorded everything said to 6-9 month old babies. Even in a situation in which mothers are asked to teach their infants new words, mothers only produce an isolated word 20% of the time. Of course, as adults, it seems very clear to us where the word boundaries are when we are listening to our own language, but if you listen to a foreign language, it is often times difficult to tell where one word ends and another begins. This is because words are run together in fluent speech. The cues that we use are particular for the language. We will first speak about low level cues, and then about higher level cues that infants may use when they get older.

One set of cues is known as stress-based cues, and they have to do with where the accents fall on words. If you have consistent stress placement with respect to word onsets or offsets, these cues could be helpful in segmenting words. A good example is a language like Czech, where stress is regularly on the first stress syllable. In Polish, the accent always falls on the second to the

last syllable, which is still helpful as a cue—but perhaps more helpful for the offset of the words. Then there are phonotactic cues, which refer to sequences of phonetic segments. There are certain clusters of phonetic segments that are common in some languages, but not in others. For example, the segment /kt/ is common in Polish but not in English. Thus, in English, /kt/ may mark a word boundary where one word ends and another begins. Another type of cue is the allophonic cue. These are contexts in which a particular phones can appear. For example, the /t/ sound in the beginning of the word TOP is aspirated, but the /t/ in POT is said to be unaspirated. If one can track where those versions appear, then that information can serve as a cue for determining word boundaries. The last set of cues are referred to as distributional cues, sometimes referred to as statistical cues, and these are co-occurrence patterns that occur across different utterances. In a sequence like HAPPY BABY, there is a relationship between the first two syllables of HAPPY—where they co-occur often, and the same thing for the two syllables of BABY. But the relationship that exists—the likelihood that /pi/ will be followed by /bei/ in a particular word—is very low in English. So this may provide listeners with a cue about word boundaries.

When and how do infants begin to segment words from fluent speech? This was a question we addressed many years ago, but we did not have the proper techniques, in part because most of the procedures we used to study infants limited us to very short stimuli. Later we developed a procedure that allowed us to present long segments of speech to infants and could measure how the infants responded to the speech. In a paper that was published in 1995 with Aslin, a method was devised to test these questions. We showed that infants 7–8 months of age can segment monosyllabic words, such as CUP and DOG. Younger infants are unable to do so. We also showed in some of our studies is that English learners seem to use the stress-based cues to locate word onsets. And they do this even when this approach results in mis-segmentation of words. Moreover, the tendency is so strong that infants can even parse a foreign language which is similar to English—e.g., Dutch.

Let us take a look at this stress-based strategy and how it may work for infants. Cutler and Carter (1987) conducted a study in which they took a sample out of conversational speech, and they looked at the patterns of words that the talkers used. They found that almost 90% of the content words in the speech sample were words that begin with a strong syllable (or a stressed syllable). Based on that discovery, Cutler and Norris proposed a metrical segmentation strategy. The idea is that, when you are listening to English speaker, you can assume that each stress is the beginning of a new word. It is a rough way of segment, but it will get you a lot of words right, except most function words and some other kinds of words. The interesting thing is that, as predicted by their model, some words (e.g., ACROSS) are more likely to mis-segmented than others, if the listeners are prescribing to the metrical segment strategy.

We wanted to know whether or infants in fact follow such a strategy. Previous work has shown that, at around nine months of age, infants are sensitive to the predominant stress-pattern of English—they tend to listen longer to word beginning with a strong syllable followed by a weak syllable,

rather than the other way around. The procedure we used is the Head Turn Preference Procedure (HPP).

In HPP, infants are presented with two types of materials, and preferences for one type or the other are indexed from a comparison of average orientation times to each type. The infant is seated on the caregiver's lap in the middle of a three-sided enclosure. There are red lights on each side of the enclosure, and a green light in the center. On a given trial, the green light flashes to get the infants attention, so that the infant stares at the center. Once the infant faces the center, an observer behind the central panel presses a button on a response button. I should mention that the observer and caregiver are wearing sound-proof earphones (the kind they use on airport runways), and the earphones are playing classical music. This is to ensure that the caregiver and observer do not hear what is going on in the experiment. Then, one of the two lights on the periphery begins to flash. Whether it's the left or right is at random. When the infant looks to the light, an audio recording of the stimulus is played. The sound will continue to play until the infant looks away for two seconds in a row, or until the trial times-out. What we are measuring is the amount of time that an infant orients to a given sound, that is, the infant's preference. Our comparisons involve how one stimulus is preferred over another.

We adapted some of the procedures of adult studies and decided to familiarize the infants with isolated instances of a pair of target words (usually produced by female talkers and spoken in a motherese style). We encouraged our talker to produce it in various ways, in terms of pitch and tone, for example. [*Prof. Jusczyk utters some examples in a diction characteristic of the speech produced by adults directed to children.*] The infants listen to each of these words for 30 s. After, in the test phase, we play infants a speech sample that contains the two target words plus other words. [*Prof. Jusczyk reads a passage aloud.*] The position of the target words was sometimes in the beginning of a sentence, sometimes in the middle, and sometimes at the end. We did not always want it in the same place. The talkers did not emphasize the target words. We would expect the child to listen longer to the passages containing words with which he or she is already familiar than with words unfamiliar. The words used for the first set of infants had the stress on the first syllable. Another set of infants were familiarized with words whose stress falls on the second syllable—e.g., GUITAR and SURPRISE. We had comparable passages from both groups.

What we found is that infants attend longer to passages that contain familiar rather than unfamiliar words, provided that the words follow the strong-weak syllable pattern. On the other hand, infants do not seem to prefer passages with familiar words if those familiar words are composed of weak-strong syllable patterns. The point is that, at 7.5 months of age, infants can segment words with strong-weak syllable patterns but cannot do the same for weak-strong syllable patterns.

How could an infant learn about the predominant stress pattern if it is not already segmenting words? One possibility is that there is an innate tendency—that we are born segmenting words with strong-weak patterns. The problem with this is that it would put infants in a bind if they were learning a language without those characteristics (e.g., French and Polish). We do not think that this explanation is plausible, but, to be honest, we do not have the

data from French infants yet. Another possibility, which we feel is much more likely, is that the bias comes from specific language input.

One of the things we noticed is that most English names and nicknames begin with a stressed syllable. Even a name like ELIZABETH, which does not begin with a stressed syllable, is shortened into BETTY, which does begin with a stressed syllable. We also know that infants are sensitive to their own names by about three to four months of age. Also, words such as MOMMY, DADDY, DOGGY, and BIRDIE also have the stress on the first syllable. So there is evidence that most of the child's inputs have the strong-weak syllable pattern.

Yet, the question remains, how do you get out of this metrical segmentation strategy? How do you segment weak-strong syllable words? The suggestion is that other contextual cues and phonotactic cues help. In fact, what our studies have shown is that infants use these higher-level cues in the subsequent months, that is, the metrical segmentation strategy is employed as a first-pass strategy, then additional cues are used. We showed that infants at nine months can use phonotactic cues when they are the only cues available. At 10.5 months of age, but not before this, infants can use allophonic cues—differentiating NIGHT RATES from NITRATES. At nine months, they cannot accomplish this. Interestingly, at 10.5 months they can also segment the weak-strong words. Another finding is that eight months old can use statistical cues that are present in the input, in the absence of any other cues. Statistical cues are based on the natural fact that some sequences, or cooccurrences, of sounds are more likely than others. We replicated this findings of others with real speech, not artificial sequences.

Now, you may ask whether these statistical cues would be sufficient by themselves, without any other cues. Elizabeth Johnson and I set a comparison pitting the metrical cues against the statistical cues. This time, in our familiarization phase, we replaced the isolated syllable with a stressed version of that syllable. Why pit the two types of cues against each other? Well, if you use the stress cues, the boundary should be different from that of the statistical cues. We found that infants rely more heavily on the speech cues.

It is pretty clear that a language learner has to rely on multiple cues—just using stress, for example, you will mis-segment words like GUITAR. What we have been doing in our lab is to try to start to understand how infants can integrate the different cues in order to segment the speech appropriately. We also want to learn how infants weigh one cue versus another. In the long run we want to know how all this information is integrated. What you begin to think is that, somehow, infants learn a strategy and then check to see if what is segmented is a plausible word in the language. Cutler has been looking at this. We started to look at the possibility that twelve-month olds may use this checking strategy and ran several experiments on the matter.

The results for several studies is that the infants (12 months old) do seem to be sensitive to the kinds of segments that could stand for possible words in their language. As adults, we use other cues. Most models of word identification use a lexical competition system, that is, to identify a word based on the best match of an utterance to a perceiver's known words. Clearly infants will not exploit such methods because they do not have much of a lexicon. By seventeen or eighteen months, infants possess a lexicon of around of fifty

words or more. We were wondering if, once an infant has a lexicon, it starts using more lexical strategies. This is work that I have been doing with a post-doc and Paul Luce. The question is, how does the sound structure of a word interact with your ability to learn a new word. A lot of the work on word acquisition has looked at the semantic factors that drive this process, but little work has looked at the sound factors involved.

The basic question here is whether or not it is easier to learn a new word that sounds like a lot of previously learned words compared to a word that sounds like few previously learned words. We could predict results in either direction. You could say that a familiar pattern is easier to learn, or that it is more likely to confuse. Part of the way of testing this is by referring to some of the insights that Luce and others have had about speech recognition. Lexical neighbors are words that differ by a single phoneme from a particular word. For example, CAT is a neighbor of HAT and SCAT, because CAT differs from them in just one phoneme.

For practical reasons, we had to create our own lexical neighborhoods with artificial words. Some neighborhoods were dense (many neighbors) and some sparse (few neighbors and many unrelated words). [*Prof. Jusczyk reads the artificial words from each neighborhood.*] Infants became familiarized with these words in six learning sessions. Then we employed the *split-screen preferential looking paradigm*. This takes advantage of the fact that children learn sound patterns even when they do not know what the sound patterns mean.

Within a four second interval, we measured how long infants looked at the pictures that were associated with words in high or low-density neighborhoods. The results are that there is no evidence that the infant learned the high-density target words, but it appears that they do learn low-density target words. The reaction time data show the same effects: infants are slower in turning to the target when it comes from a high-density neighborhood. So infants appear to be affected by neighborhood density effects. High-density items, words that sound like many other words, are more difficult to learn. In adults, in general, the same thing happens, though you get facilitation from non-words, in contrast to real words.

I will attempt to sum it up here. English learners begin to segment words around 7.5 months of age, using low level speech cues such as prosodic stress. They begin integrating different cues by one year of age. Somewhere around seventeen months, you begin to see lexical competition effects, which means that, at this stage, infants are using low level and lexical level cues to serve the development of lexical knowledge.

## APPLAUSE

### Questions from the Audience

**Prof. Krauss:** I have a really naïve question. Why does the child look in the direction of the familiar words?

**Prof. Jusczyk:** We purposely had the speakers produce the stimuli in diverse ways, and this would get the child's attention. If you look in our passages, there are many words other than the target words, so there is a lot of variety. Under these circumstances, habituation is unlikely. What the infant is trying to

do is discover what is going on and how the language is working. In this context, the infant is looking for regularly occurring patterns. In our lab, we used this familiarity paradigm over 150 times, and, amazingly, we always found the same thing, even though we use different stimuli. Rarely did we observe novelty effects, however, as in the Saffran paradigm.

**Prof. Krauss:** Do you think that having different speakers will have the same effects as having the same speaker utter the words in different ways?

**Prof. Jusczyk:** We have done experiments in which we used different speakers, especially in the word-segmentation experiments. It raises interesting questions: are you storing a specific exemplar or are you storing something more abstract? Many experiments have shown that talker characteristics are important, and that the representations are stored in reference to the voices. You can get infants to generalize between two different female voices or two different male voices, but they do not generalize to another sex until 10.5 months of age. The infants seem to encode speaker information.

**Prof. Krauss:** Over what criteria do infants generalize to other voices of the same sex?

**Prof. Jusczyk:** Well, we have not broken it down that far, but one of my students has done a dissertation on that. It turns out that the similarity space of the voices influences generalization processes.

**Prof. Rosman:** Why did you employ multimodal methods instead of just auditory?

**Prof. Jusczyk:** The multimodal presentation was motivated by the question addressed earlier of whether or not the actual sound matters, or whether the representation is more abstract. We could use the false memory paradigm to see whether what matters is the actual memory file, or whether what matters is some abstract rendition of it.

**Prof. Van Lancker:** In speaking about stress and stress sequences, what role do vocatives play as an input, and how do they affect segmentation strategies?

**Prof. Jusczyk:** Some of those vocatives would be like HELLO and GOODBYE, so it is a little bit harder to say. To be honest with you, I believe that the work of van de Weijer should be published because it is an enormous corpus that has *all* the things that were said to the infant—very valuable indeed.

**Prof. Krauss:** But there is no control condition!

*[Laughter from the audience.]*

**Prof. Jusczyk:** Right! I was waiting for someone to ask this question...

**Ms. Piorkowski:** Actually, I was wondering, have you ever tested the linguistic capacities of cochlear implant infants in terms of speech perception?

**Prof. Jusczyk:** Actually, you were supposed to ask about how the neighborhood stuff works out. *[Laughter in the audience.]* No, I have not, but someone in our department has been trying to get some perceptual measures from an infant 18 months of age.

**Prof. Van Lancker:** How about bilingual children?

**Prof. Jusczyk:** Unfortunately, all of our experiments were done with English speakers from English-speaking households. Interestingly, the largest immigrant group to Baltimore is Korean, and I have yet to find an expert.

Some experiments are being conducted now with French-English bilinguals in Canada. We have been doing some cross-linguistic studies. Mehler has been looking at bilinguals.

**APPLAUSE**

---

**Place:** Faculty House  
Columbia University  
400 West 117th Street

**Time:** 4:00 PM

**Chair:** Prof. Robert E. Remez, Barnard College, Columbia University.

**Attendees:** Mary Anne De Fuccio, Aili Flint, Peter Gordon, Liz Henly, Robert M. Krauss, Diane Masropieri, Bruce McCandiss, Michele Miozzo, Ezequiel Morsella, Katherine Nelson, Rebecca Piorkowski, Lois Putnam, Robert E. Remez, Abe Rosman, Paula Rubel, Ann Senghas, John J. Sidtis, Diana Van Lancker.

**Rapporteur:** Ezequiel Morsella.





22 MARCH 2001

---

## Universal and evolutionary aspects of cross-language color naming

Paul Kay  
Department of Linguistics  
University of California, Berkeley

Throughout the 1940s, 50s, and much of the 60s the Whorfian doctrine of radical linguistic relativity was dominant in the social sciences. The *locus classicus* of this doctrine was the lexical domain of color words. A typical pronouncement of the period was, "Our partitioning of the spectrum consists of the arbitrary imposition of a category system upon a continuous physical domain" (Krauss, 1968). Berlin and Kay (1969) challenged this view on the basis of an experimental study of color naming in 20 languages and reanalysis of the accounts of the basic color nomenclatures of 78 additional languages found in the linguistic and ethnographic literature. That study found universal constraints on possible color naming systems as well as evidence for a partial evolutionary ordering among the limited range of color systems discovered. It concluded that the explanation for the particular patterns found must lie to considerable degree in pan-human processes of color perception and representation but confessed inability to produce such explanations. Research conducted over the past 30 years has made advances both in sharpening the empirical claims of the original study and in relating universal constraints on color naming systems and recurrent patterns of historical development of these systems to independently established facts regarding color appearance.

In my graduate school years, everything was about linguistic *relativism*, and color coding was a prime example of it. It is interesting that, although color perception and naming were central issues in debates about linguistic relativism, Sapir and Whorf never spoke about color coding.

Around 1965, my friend and I had the same experience while studying the color names of two languages. He studied a Mexican language, and I studied a language from Tahiti. Nevertheless, we were surprised that the color names of both languages translated well into English, except for having just one color term for green and blue. (What a coincidence that both languages, from such foreign regions, shared this anomaly!) We then wanted to test whether or not every language mapped color names onto perceptual color space. Our procedure was the same as that used by Lenneberg and Roberts. The first thing we did was to elicit color terms in the absence of color stimuli. Now, this idea of a "basic color term" is contentious and not uncontroversial. The definition of a basic color term is: *The smallest set of words with pure color meaning with which a speaker can name any color.*

After eliciting the basic color terms, we asked subjects to perform two tasks. In the first task, subjects were asked to point out the best example of a given color from a sample of colors—again, we used the color charts used by Lenneberg and Roberts.

*[Prof. Kay projects a color chart on the screen. The display consists of a graded set of color chips spanning the spectrum, from red through blue, and from dark to light.]*

The second task was to select all the samples that describe a given color.

The results of our experiments can be explained by two terms: evolution and universals.

*[Prof. Kay projects a version of the color chart on which regions of chips have been inscribed. Prof. Kay describes the charts.]*

Each dot upon this chart shows the best example for each color, for each of the tested languages ( $n = 20$ ). The number above each dot shows the number of times that each cluster was chosen as containing the hue of the color name. There were five languages that had six color terms. Overall, we found that the between language differences were no greater than the within language differences. This supports the idea of universality.

We then went to the dictionary and analyzed reports from the literature which included another seventy-eight languages—so we had a total of ninety-eight languages.

The conclusion that one must draw from these data is that color naming follows an evolution, an evolution that is guided by principles and follows one of several trajectories. The evolution model begins with a “white/black” distinction which then can be followed by a “red” distinction, and then you may have a color term for blue *and* green, and then later the color terms distinguish “blue/green.” These were the simple beginnings. Although there have been changes at the level of detail, the Evolution Model of color naming still holds true.

Since then there has been a succession of models motivated by two considerations: 1) more linguistic data, and 2) more knowledge of color perception. With regard to color perception, at the end of our original monograph we wrote, “a finding of universals in color naming must be due to the way we perceive and represent color.” The perceptual component of this explanation refers to the opponent-processing color theory, which shows that within a basic set of six colors, there are dimensions of graded variation and opponency. Because of the opponency process, we have no sensation—or idea of—a color sensation such as “blue–yellow” or “red–green.”

Linguistic categories follow primary colors. There are primary colors and then secondary and tertiary colors (e.g., orange, purple, gray, pink, brown, etc.).

Now we test the *emergence* hypothesis—the presupposition that languages have a set of words to describe the basic color experiences. We refer to the World Color Survey, in part compiled through the work of linguistic missionaries. If we look at the data, obtained from 25 speakers from each language, we see that this assumption is challenged: some researches (e.g., Levinson) have found that some languages have color names for just white, black, or red, and my collaborators and I have found additional cases.

Four certain principles dictate the evolution of the generation of color names. The first principle is the sociolinguistic concept of *partition*: that a language will assign meanings to words to carve up a denotative domain. More directly, a language will use a set of words to analogous to the breaks within a salient domain of experience. This principle explains how a language without a color term will engender a color name when a communicable representation of the experiential contrast becomes useful. And then we have three principles based on color *appearance*. The first is to distinguish black/white, based on the perceptual fact that with variation in black and white alone we can have good object recognition, without hue. This is also supported by physiological evidence about the parvo- and magno-cellular levels of the lateral geniculate of the thalamus. The second appearance principle is to distinguish the two classes of warm and cool colors, which is not motivated by strong evidence but has its theoretical roots all the way back in Aristotle. Warm colors are red, orange, and yellow; and cool colors are green and blue. A third appearance principle is to distinguish red, buttressed by some theories about the wavelengths of the hue and on the linguistic fact that red is usually the first color name in a language.

Of the World Color Survey data, the evolution of all color names in all but six of the world's languages follows the principles mentioned above. That is, one can predict the trajectory of the evolution of color names. For example, if there are only two color names, they will be for white and black. Some trajectories are more likely than others.

*[Prof. Kay shows a slide in which the color name history of each language is represented as a different path, describing the points of common origin and orderly points of divergence.]*

When we apply the principles in logical order, we account for 92% of the languages.

The current model with the four principles (one sociolinguistic and the other three based on color appearance) accounts well for the international data. Within these parameters, color naming appears to be well determined.

## APPLAUSE

### Questions

**Prof. Krauss:** The motivation for your quotation was to try to illuminate the relation between language and cognitive processing. There are few people who believe that color coding is the best arena to test the relation between language and thought—the war was already fought and won. I think the important thing is to study the relationship between cognition and *language use*, an area of study that has been quite fruitful.

**Prof. Kay:** I agree—by using color, the relativists really used poor judgment, for we knew *then* that there are different receptors which are sensitive to different wavelengths. I certainly agree that one cannot generalize from these findings to phenomena pertaining to the effects of language use, and that the Sapir-Whorf hypothesis cannot be shot down with the color coding data. Some of Whorf's writings convey that he was not as much a relativist as most people

think, or perhaps not a relativist at all. The idea that language can impose a representation upon the world, and that these representations can affect nonlinguistic behavior, carries many assumptions. I think that the first assumption is wrong: there are not too many ways that language can speak about space. However, the way this is done can affect cognition.

**Question. Prof. Gasparov:** In color naming, what do you find with the borderline colors, e.g., where red ends and yellow begins.

**Prof. Kay:** We have not tried to establish how much variation there is amongst languages when speaking about boundary conditions as you mentioned. It is hard to know by looking at our data how much variability is due to languages and how much is due to the speakers.

**Question. Prof. Remez:** English preserves the terms warm and cool for colors that also have unique names. Do other languages preserve the older, less differentiated terms once the innovation of more differentiated terms has occurred? And, are warm and cool pure terms of color or are they metaphors?

**Prof. Kay:** This is a difficult issue because of the availability of pure color terms for metaphor. To give an example, my wife said, "This lemon is green." Did she mean "green" or "unripe." The answer is that there is no single answer—the semantics of polysemy plays an important role.

**Question. Prof. Terrace:** Is there any data on the names to describe kin that show any universality?

**Prof. Kay:** Yes, the most important parameter is lineality; the second is generation. There are a few patterns that over-ride these distinctions by mixing lineality and generation.

**Question. Prof. Terrace:** Do they follow a trajectory similar to that of color names?

**Prof. Kay:** Systems of pronouns and numbers do not always work the same, but there are some generalities that come up with different languages. Then again, after about the first set of one-hundred words, you still have the rest of the lexicon to worry about, so there is some variation.

**Question. Prof. Miozzo:** Your approach is rather indirect: a more direct approach is to look at a cluster of related languages and to see whether the variations in them is predicted by your model.

**Prof. Kay:** That question is best answered by historical linguistics. In the literature on Latin, for example, there is evidence that there are Italic languages today that do not have a separate term for blue. One quite perplexing thing is that Ancient Greek had a blue-green term and that Latin had a green-yellow term. Then again, as soon as you get to proto-languages, the arguments about reconstruction are circular.

---

**Place:** Faculty House  
Columbia University  
400 West 117th Street  
**Time:** 4:00 PM

**Chair:** Prof. Robert E. Remez, Barnard College, Columbia University.

**Attendees:** Lila Braine, Norman Ferrer, Aili Flint, Boris Gasparov, Hui-Chen Hsu, Marco Jacquemet, Dustin Merritt, Mobina J. Khan, Johnne Kleifgen, Henry Kong, Lauren Kornrick, Ilona Kovary, Robert M. Krauss, Dustin Merritt, Michele Miozzo, Ezequiel Morsella, Tomislav Pavlicic, Rebecca Piorkowski, Robert E. Remez, Jay Sandhaus, Penny Shima, Simon Teufel, Michael Torpey.

**Rapporteur:** Ezequiel Morsella.





3 MAY 2001

---

**Language acquisition and intentionality:  
the essential tension between  
engagement and effort**

**Lois Bloom**

*Department of Psychology and Human Development  
Teachers College, Columbia University*

Theories of language acquisition have become increasingly mechanistic and deterministic, and the result is the objectification of both the child and the language that the child is acquiring. Words and sentences, as objects of study, often assume a life of their own, apart from the child who produced them and apart from the situations in which the child said them. But when the units of language are removed from the very fabric of the child's life in which they are necessarily embedded, separated from the so-called extraneous variables of performance, then the language the child is learning becomes disembodied and decontextualized. Giving the child's language acquisition both an embodiment and a context is my primary goal in this presentation. The dominant themes in my program of research are the intentionality of the young child, as the agent of the acquisition process, and the essential tension between engagement and effort that propels the process forward. Behaviors depend on one's interest and engagement in an event as well as the attention and effort that different actions and interactions require. When we have to do two or more things either simultaneously or successively, performance reflects accommodation to the effects of task difficulty. The research I will describe examines the intricate and complex adjustments in the convergence of multiple behaviors that are coextensive with language in everyday events and the coordination of their development over time. Two conclusions follow from these studies. (1) Language is not separate, but is acquired as part of a child's development more generally. (2) Acquiring language isn't easy, and performance counts.

It is fitting that I give my last lecture at Columbia, where my career began under the sponsorship of William Labov, who was on my dissertation committee. My first public lecture was a University Seminar, much like this one, and it was a real baptism of fire! A very formative experience. The entire audience was male except for this one woman whose theory was entirely at odds with mine. My theory dealt with the acquisition of word meaning, at a time when everything was about acquisition of forms. My back was against the wall! I began this research twenty years ago, and my interest has been how language acquisition affects other aspects of development. Fortunately, during all these years I have always been accompanied by a very talented group of

students. Much of what will be presented today will appear in an upcoming monograph.

I will start out by giving you some background. I will dwell primarily on three themes: 1) the child who acquires language, 2) theories of acquisition, and 3) the idea and importance of *tension*, both for the development of the child and of theory.

In the last twenty years I have been concerned by the fact that the child has been seen as an object, not an agent. I call this the objectification of the child. Also, language has been looked at as an object, that is, words, sentences, and nouns are things removed from the speaker. When science does this, language becomes disembodied and decontextualized, and the child becomes the repository of skills and features, which together make language possible. I have been trying to see things in an opposite way, by showing how language is contextualized within the development of the child. I will begin with theory.

As Thomas Kuhn elucidated, theories experience tension when they can no longer accommodate observations. This tension is necessary for theoretical development. What I will argue is that theories of language acquisition emerged from a succession of psycholinguistic tensions. These tensions characterize the theories that have been proposed. I will present a simplified historical survey.

Tensions arose in the late 1950's when B. F. Skinner's *Verbal Behavior* (1957) was challenged by Noam Chomsky in 1959. *Verbal Behavior* talked about "other people"—what they did or said determined how language is acquired. Is that fair Herb Terrace?

**Prof. Terrace:** Yes it is. [*Laughter in the audience.*]

**Prof. Bloom:** This behavioristic interpretation of language was the dominant view at the time. In contrast, in a review of *Verbal Behavior*, Chomsky claimed that children do not learn behaviors but a grammar that does not exist in the environment. As Chomsky persuasively set forth, if it does not exist in the environment, it must exist in the child's head, and thus be innate. This was the first tension: environmental theories (two-person theories) versus one-person theories (what is in the child's head). This is also the beginning of language-acquisition research in the last century (we can now say "the last century," wow!), and the beginning of the objectification of both the child and language.

Later, my research and that of others, as Roger Brown pointed out, showed that if you are simply listening to the form of language and not the meaning, you are missing quite a lot. What it is that children learn, when they acquire language, is how to express meaning, and meaning in language comes from what we know about the world. This new tension was between cognition and language. Keep in mind that it is still a "one-person" theory, and it is still an objectified child that learns language. It wasn't very long before people pointed out that it was not enough to have a language-acquisition-device (LAD), but that certain social processes had to occur for successful language acquisition. The social aspect of language acquisition was put forth by several theorists, including Robert M. Krauss. Now began a two-person theory, a tension which continues today between the social and individual aspects of acquisition.

More recently, in the late 1970's and early 1980's, another psycholinguistic tension arose in the invocation of linguistic theory. Children's language phenomena served as evidence for or against different linguistic theories. But again, these theories, like those of Wexler and Pinker, also treated language as an object, as decontextualized. It all goes back to Chomsky's LAD. Then another tension arose from connectionist theories such as those of Rumelhart and McClelland. From this point of view, children are, not learning rules or a grammar, but acquiring neural connections from linguistic input, connectivities that enable them to use language. In short, these theories look at the biological substrate of language. This tension is still alive today and is exemplified by the conflict between MacWhinney at Carnegie Mellon and Pinker at MIT. Again, both the language and child are objectified. The result is a child who is a repository for different mechanisms that impinge on and determine language acquisition.

What is missing in all of these theories is the *intentionality* of the child, that is, the child's agency, the role that the child is playing in language acquisition. Once you embrace that it is the child that drives all the tensions, then you see that the child is the subject, and language becomes the expression of the child's intentionality. Language *can* be looked at as an object; however, a language is never acquired without a context in which engagement and tension are important. The child's engagement and effort is required. Language requires cognitive resources, and the child's resources are limited and have to be shared with other aspects of development. Language is acquired within a social and emotional context where these limited resources have to be shared. Thus, development depends on tension.

Most theories of development refer to the notion of tension. The idea of tension can be seen in the work of Freud. Freud's theory is based on the notion of tension reduction, having to overcome the pleasure principle; the child could never develop by just reflex. Piaget's *equilibrium theory*, though not espoused as such, also dealt with tension reduction. There is the need to maintain equilibrium in the face of new inputs that resist understanding, for instance, through the device of accommodation.

Language acquisition likewise depends on an essential interplay between *engagement* and *effort* in the young child's intentionality. I am taking consciousness and intentionality very seriously. (This theory was influenced by thinkers here at Columbia, people like Charles Taylor.) Words and sentences are embodiments of intentional states and representations, and language is learned as an act of intention. Children learn words and sentences used for expression and set up *by* interpretation. Children's language serves to express and to articulate the elements, roles, and relations represented in their intentional states, and these states are needed in order to interpret what other people say, that is, by attributing intentional states to the speaker. Moreover, language is learned in this process. Children are acquiring language in acts of expression and acts of interpretation.

The Intentionality Model of language acquisition has the components of engagement and effort, and there are three important principles: 1) relevance, 2) discrepancy, and 3) elaboration. These are broad principles to explain the role of intentionality in language acquisition, and they capture the role of

intentional states and their interaction with things in the outside world. The principle of relevance states that development is enhanced when events are pertinent to what the child has in mind and are worth knowing. The principle of discrepancy states that development is enhanced when the child acts to resolve a mismatch between what the child and other persons have in mind, or/and between what the child already knows and the knowledge produced by new encounters that resist understanding. And last, the principle of elaboration states that children will have to learn more of a language in order to keep up with developments in other areas of cognition. The representations even at the first year of life are too complex to convey through expressive behaviors. These principles mediate between engagement and effort.

In the last twenty years we attempted to test a number of these ideas. We looked at developments in language in relationship to other aspects of development. In three studies, the relation between language development and the development of other cognitive/emotional processes was examined. We studied twelve children from nine months of age to the age of two years. We saw them once a month during which visits with their mothers they played with toys. The primary data deals with the natural behaviors of the children in this playroom setting. The observed changes over time were attributed to the child because the objects and context of the situation remained the same.

One of the few things which everyone in the field agrees upon is that, in language development, there are three important landmarks: the first word in the laboratory (mean time = 13 months, range = 10 to 17 months), the vocabulary spurt (mean time = 19 months, range = 13 to 25 months), and the simple sentence (mean = 24 months, range = 18–32 months). As you can see by the ranges, there is great variation. We look at development in reference to these landmarks. We also did studies which looked at development over chronological time, but I will not be speaking about them today.

The first question dealt with the relationship between systems of expression early one of affect and later ones of language. In comparison to expressions of affect, such as crying and smiling, early language is poor at communicating. The child is quite effective at conveying emotions through expressive behaviors, and caregivers are adept at receiving and interpreting them. We are going to look at these expressions as evidence of engagement. In contrast, early words and sentences are fragile and imprecise. Expressions of affect are unlearned and quite functional, whereas language has to be learned and requires effort. We looked at behaviors in real time and had to find a way of separating the relevant behaviors. I will now explain how the data for our study were collected.

*[Prof. Bloom shows a graph depicting the experimental set up of 1983. The children and their mothers were recorded in a play area. In a different room, various workers coded different behaviors that were sampled on audio and videotape. Different workers coded different things: some coded speech and some coded visible gestures. All recording apparatus were synchronized, and all recordings possessed a time stamp. An Apple 2 Plus controlled the data collection procedure. With the time stamp, the number of video frames in which a behavior occurred could be counted and measured. All the different*

*recordings, audio and video, could then be patched back together, and the temporal relationship amongst the different coded behaviors could be examined.]*

One of the earliest questions that we asked was, What percent of video frames contain child speech? We discovered several things. First, children talk more over time, but the amount of emotional expression, in relation to landmarks, interestingly, stays the same. Thus, this suggests that the development of language does not represent an overall increase in expressivity. Also, language does not replace emotional expression. Second, early (6 children) and late word learners (7 children) did not differ in how much time they spent talking (even though they are different in chronological age), but they do differ in emotional expression: later word learners spend more time expressing emotion than early word learners. Again, these developments occur over what I call developmental time.

The other studies deal with the occurrence of behaviors microgenetically. We used a lag segmental analysis, in which the temporal relation between expressive and language behaviors was examined over immediate real time. We looked at the occurrence of emotional expression during, before, and after children said words and sentences. We also looked at the relation between these two forms of expressions and play.

In the first slides, the target behavior will be speech. A second behavior will be lagged to the target behavior (speech) in 1 s intervals.

*[Prof. Bloom shows a graph depicting the occurrence of different behaviors over time.]*

The first hypothesis was a hypothesis of effort, based on a model of limited resources. Kahneman's theory of 1973 states that there is a limitation on the amount, quantity, or capacity of resources that can be allotted to one task at any given time. What the hypothesis of effort says is that, if multiple tasks tap into the same single pool of resources, then performance will suffer. Adapting this view, the idea here is to see whether emotional expression, play, and language share the same resources, or the same intentional state. If so they should compete for resources. Thus, there should be greater effort at times of emergence of the first words than at times of achievement, for instance, at the vocabulary spurt. These developments will tap the limited resources of the child. Engagement would lead to a discrepancy, and overcoming this discrepancy will require effort. Emotional expression, as an index of engagement, you would expect more emotional expressions at time of achievement in new learning rather than at time of emergence and transition. But because emotional expressions induce arousal, you will get bi-directional effects. Again, Kahneman's idea is that arousal should hurt learning, because of the limited resources. This idea, also incorporating the notion that emotional expressions require cognitive resources (because they are appraisals of situation), has recurred frequently in different theories (e.g., those of Mandler). If that is the case, then there should be bi-directional effects. Late word learner may show more emotion because the arousal from emotional expression may interfere with language learning. Or perhaps children who express less emotions can suppress arousal to establish a context for learning.

The first result of our study (Bloom and Beckwith, 1989) is that, at the first-words milestone (a time of transition, an emergence), there is a decrease in emotional expression before the utterance. We attribute this to the arousal during the formation of the intentional state and during the recall of the appropriate word, which occurs seconds before the utterance.

**Question from Audience: Prof. Terrace:** Is the mother present, and did you code whether the expressions were directed to her or not?

**Prof. Bloom:** Yes, but what is coded is just the onset and offset of the expression. (To whom the emotions and language is directed is in the coding system, but I won't be referring to that data today.) During vocabulary-spurt (a time of achievement), the children has got the two systems of expressions together—children could talk and express emotions at the same time.

**Question from Audience: Prof. Remez:** Do all aspects of emotional expression rise and fall together?

**Prof. Bloom:** No, there were certain constraints, for we only coded certain behavioral cues, such as facial expression, posture, and vocal tone, independent of speech.

When we looked at the words that occurred during emotional expressions, we found that they were either the children's most frequently used words or the words that were learned earliest. They were not infrequent or recently learned words. When we looked at the emotions that were expressed, more positive affect accompanied words than negative words. We interpreted that to mean that more effort is required for negative emotions, which is consistent with cognitive-emotions theory. Theory claims that negative emotion is an obstacle to a goal. Sadness and anger have to do with a goal is lost or blocked, respectively, and positive emotions occur when a goal is achieved. So what happens during the transition to multi-word speech? We did the same analyses, and found that emotion is dramatically suppressed. This means that at the time of transition, children are likely to express neutral affect. We looked at early and late multiple-word learners, and found that early learners now show more emotions during speech than late learners.

Then we looked at two kinds of speech: imitated and rehearsed speech.

Overall, the transition to sentences requires work, and this work reduces the tendency to express emotions during speech. Moreover, late learners show that effect more than early learners. Late learners suppress emotions more and are hypothesized to have a harder time making the transition. Also, imitated speech is easier and children can express more emotional behaviors during this time.

All of these studies looked at speech and emotion as they co-occurred. These studies used all language events as targets, regardless of what was going on in the context. We know that a lot was going on—e.g., mothers were having snacks or talking to investigators. We then looked at a subset of these events, especially play. We found that at nine, ten, and eleven months of age, children were not putting parts of toys together. They were taking things apart, for example, by knocking-over and dumping things. Their mothers would put the things back together again. What we found is that the time at which children began to put the objects together coincided with their transition to first-words.

The next analysis will show you what happened at the vocabulary-spurt in reference to a different target event (putting objects together in a thematic relationship). We lagged to this target event to two kinds of behaviors, speech or emotional expression. In the moments around constructing activities, children tended not to express emotions. We are interpreting this as the cognitive demand of emotional expression and play. Of course there are other hypotheses. In the moments of construction, the children are more likely to express themselves with speech than with emotion. There is a trade-off between these two kinds of expression. Another target event is child-and-mother speech interaction. Overwhelmingly, children are more likely to be talking before the mother speaks and least likely to be talking while the mother speaks. Then there is an increase after the mother speaks. Mothers are more likely to talk after the child speaks. Combining this with emotional expression, we find that mothers are least likely to be talking during play (constructing activity) and more likely to talk after the constructive play activity. Just as the mothers were primarily responsive to their children in conversation, overall they are also responsive to children during their activities.

My three conclusions are the following. Despite what you heard from MIT, acquiring language is effortful and children are working at it. It isn't easy. Also, despite what you heard from MIT, performance does count. Chomsky's entire theory has been predicated on a competence mechanism that has served to denigrate performance, but, in fact, performance counts. Children are learning language in acts of intention and expression. Finally, language is not separate from other capacities. It is not a separate module; language is acquired in relation to the rest of the child's development, which leads me to the following mantra:

"Every behavioral act, whether outward bodily movement or internalized cognitive operation, gains its significance and status in terms of the overall functioning of the organism" – Heinz Werner, 1963

## APPLAUSE

### Questions

**Prof. Terrace:** What can you tell us about pre-linguistic communication, and how it is related to intentionality? Is it the emergence of intentionality and the child's knowledge of the intentionality of others that drives linguistic development?

**Prof. Bloom:** We started at around eight to nine months of age. We have not looked at behaviors occurring before this time. In respect to current theories on intentionality, first of all, they focus on the intentionality of the caregiver. The theory is that the child cannot utter the first word until he or she has recognized the intentionality of others, but this theory denies the reality of the representations that the child has virtually from birth. The difference is that, early on, the infant's representations are determined by perceptual experience. Pretty much what the infant has in mind has to do with the immediate perceptual state. One of my favorite things is to watch six-month-olds—they

are not mindless. They are looking around and learning about their environment. Later, previous experiences have a greater role in the intentional representations. At what time does the child have a conscious appraisal of the intentional states of others? That's a deep question.

**Prof. Krauss:** First of all, I would like to say that this was a wonderful talk. As you know, in the literature on emotional expressions, some theories state that facial expressions are automatic, and some state that they are an intentional responses, as communicatively intended as words.

**Prof. Bloom:** As the philosopher John Searle pointed out, there are different meanings of the word "intentionality." I am using the term "intentionality" with a capital "I," as John Searle does. Basically, it means that something is an expression of mental content, not that the child *intends* to smile or cry. The child possesses an intentional state.

**Prof. Metcalfe:** If someone were concentrating, is that emotional? What does the mother do during the child's positive state after construction? And in the speech interaction, is the mother simply echoing what the child uttered?

**Prof. Bloom:** A student of mine looked at that in 1989. What she found was that, typically, the mother expresses emotions (though not the same as that experienced by the child—e.g., when the child expresses negative emotions, the mother can express positive emotions) or explain the causes of the child's emotional state.

*[Prof. Bloom then shows a short video of a child demonstrating the three different emotional states: positive, neutral, and negative.]*

We coded emotions continuously. We also coded different levels of intensity.

*[Prof. Bloom shows a short video of a child demonstrating the three levels of intensity: slight, mild, and full.]*

Interestingly, children display neutral affect 84% of the time. I should mention that a remarkable fact of this study was that, in all the years, we only missed one session with one subject!

**Prof. Krauss:** Did you code everything frame-by-frame?

**Prof. Bloom:** Yes.

**Prof. Krauss:** Goodness gracious!

**Prof. Miozzo:** You emphasize limited capacity, but some evidence that you presented goes against this: children talk while playing. This makes language appear as special.

**Prof. Bloom:** Our interpretation is that the suppression occurs during language planning, so the evidence is not contradictory.

**Prof. Metcalfe:** During play, what does the kid say? Is he or she merely translating the motor programs?

**Prof. Bloom:** They are saying things about the objects.

**Prof. Terrace:** Are they talking about perception?

**Prof. Bloom:** About everything that he or she is doing.

**Prof. Senghas:** I am addressing your first point that language is not easy, at least in comparison to emotional expression. What do you mean by “not easy”?

**Prof. Bloom:** It is not easy because it interferes with other cognitive processes such as emotional expression.

**Prof. Remez:** Are emotional expressions also imitated proficiently in the imitation task?

**Prof. Bloom:** The easy answer is that we have not looked at that, but I would say that the mother and child do not share the same expressions.

**Prof. Putnam:** Would you find more negative emotions with less frequent words?

**Prof. Bloom:** Infrequent words are accompanied by neutral emotions.

**Prof. Bloom:** Thank you all very much.

---

**Place:** Faculty House  
Columbia University  
400 West 117th Street  
**Time:** 4:00 PM

**Chair:** Prof. Robert E. Remez, Barnard College, Columbia University.

**Attendees:** Lila Braine, Patricia Brooks, Gerald Echterhoff, Gisela Jia, Marco Jacquemet, Ilona Kovary, Robert M. Krauss, Christoph Mensebach, Janet Metcalfe, Michele Miozzo, Ezequiel Morsella, Rebecca Piorowski, Lois Putnam, Robert E. Remez, John Saxman, Ann Senghas, John J. Sidtis, Herbert Terrace, Diana Van Lancker.

**Rapporteur:** Ezequiel Morsella.

