In L. McNair, K. Singer, L. Dolbrin and M. Aucon (eds.),
Papers from the Parasession on Theory and Data in Linguistics
Chicago Linguistic Society 32:77-106, 1996

When Intuitions Fail

William Labov, University of Pennsylvania

The title of this article looks in two different directions. It asks, "When do intuitions fail?" hoping to locate the conditions under which intuitions are likely to beunreliable.But it can also be the subordinate clause, of the question, "When intuitions fail, what can we do?": Both approaches, the remedial and the prophylactic, will be considered here.

In 1951, Carl Voegelin and Zellig Harris outlined the basic opposition between two methods of gathering data. They characterized the method of Boas and Sapir as simply "Ask the informant," and pointed out the many biases introduced into the data by this method. They acknowledged that the second method, to record the speech of the informant directly, also involved many biases, particularly in the artifacts introduced by having the speaker wait while the investigator transcribed the data. They were hopeful that the recent invention of the magnetic tape recorder would resolve these difficulties, and allow linguistics to move forward on a more objective and scientific basis.

My own bias is on the side of Voegelin and Harris, towards the study of speech behavior. However, it must be acknowledged that linguistic analysis will always rely to a large extent on elicited judgments, the intuitions of the native speakers. Though studies of the speech community have focused primarily on spontaneous speech, they have not neglected the subjective dimension. Investigations of regional grammatical forms and cross-ethnic differences have searched for answers to six types of questions to assess the linguistic knowledge of a given rule or construction (adapted from Labov 1972)

X = the form or construction in question
WXY = a sentence employing X.

(1) Recognition of the form as a proper part of the language in question. ("Is WXY English?")
(2) The ability to identify the regional, social or stylistic level ("What kind of English is WXY? Who would say that?")
(3) The ability to interpret the meaning of the form in a sentential context. ("What does the sentence WXY mean?")
(4) The ability to interpret the meaning of the form out of context. ("What does X mean in WXY?")
(5) The ability to predict the acceptability of X in other contexts. ("If you can say WXY, can you say WXZ? can you say VXY?")
(6) The use of X in spontaneous speech with native frequency and native pattern of categorical and variable constraints.

The methods used to gather these six types of information are quite varied, and the methodological problems are quite different. The first five questions fall into the domain of elicited judgments: efforts to get at the linguistic intuitions of the subject, which will be the main topic of this paper. The sixth question involves the wide range of methods used in sociolinguistic studies within the community: defining and sampling the community, recording spontaneous speech across a range of contextual styles, reducing the effects of observation, defining the linguistic variable and variable constraints upon its frequency, quantitative measures of the fit of theoretical model to observation and significance of the results. These are the data that in the long run will allow us to assess the validity of the answers to the first six questions. However, many linguistic phenomena may never be studied with this degree of accuracy for several reasons:

- The time and effort required for the analysis of the use of linguistic forms in speech means that the number of such inquiries is limited
- Many linguistic features of interest are of such low frequency that quantitative studies are not feasible
- For many features the principle of accountability is inaccessible: it may not be possible to close the set that defines the variable, and say when the feature does not appear.

This last point is an overwhelming factor when we are dealing with little-known languages, where the knowledge base needed to define such a variable simply does not exist. It will therefore be necessary to pay considerable attention to the methods for posing questions (1-5). These are distinct questions, though in one form or another, all have been included in the various discussions of judgments of acceptability or grammaticality.. The usual conception of "judgment of acceptability", expressed as "Can you say this?" is equivalent to question (1). Thus in studies of positive *anymore* sentences, subjects are asked questions like "Can you say 'Farmers are pretty scarce around here *anymore*?'" If the scale given is a binary one, this is equivalent to question (1). If the scale has many values, it may include responses such as "I don't say this, but others do", or "I've heard this from some speakers from ___, though it isn't used in my community", or "Someone might say it." This is equivalent to question (2). Many students of grammaticality prefer to ask about meaning wherever possible, rather than acceptability, entering the domain of questions (3) and (4). Thus inquiries into positive *anymore* pose the question, "What would it mean to say *He's smoking a lot anymore*?" or provide a choice of *still, nowadays*, or *not at all*. Question (5) resolves into a series of questions of the type (1), but related to each other in a way that implies that some of them are English sentences, and some are not.

Finally, we note that all questions on acceptability of type (1) or (5) can be posed as relative acceptability, "Which do you like better, A or B?" and many such studies show higher degrees of agreement than an absolute inquiry into the acceptability of A. This may allow linguists to exercise an already-formed scale of acceptability or grammaticality that include the notion of "crashingly ungrammatical" as opposed to merely "ungrammatical". But for a naive judge, it has the defect of suggesting that both A and B are possible forms of the language. Some of the most striking results of grammatical inquiries occur when many judges agree that a certain form is completely unacceptable, yet use it themselves freely in every-day speech.

In several previous publications, I have expressed some skepticism about the reliability and validity of introspective judgments, particularly when they are produced by the same person that is producing the theory (Labov 1975). The most common type of response is to assert that there is no serious problem (Newmeyer 1983). Two substantial publications have appeared recently which confront the problem directly, arguing for more careful, empirically grounded methods that follow the accepted canons of scientific evidence.

Bard, Robertson and Sorace 1996 demonstrate that techniques drawn from psychophysics can give us more reliable judgments of grammaticality that agree in general with theoretical predictions: in fact, fine-grained linear displays of relative acceptability of sentence types. The authors point to "the inherent inadequacy of the measuring instrument used for linguistic acceptability judgments" to motivate their introduction of the technique of magnitude estimation (p. 34). They demonstrate that magnitude estimation of acceptability can be used to distinguish natives from near-natives whose speech and writing were virtually indistinguishable from natives. Their experiments also provide strong confirmation form naive native speakers of the linguists' judgments on the use of auxiliary *essere* or *avere* with a range of unergative, unaccusative and restructuring verbs

Schütze (1966) provides a thoughtful and penetrating review of all of the issues involved in re-designing the empirical base of syntax, from a position within the generative paradigm. Both Bard et al. and Schütze agree that it is no longer possible to set aside the analysis of variable syntactic judgments in favor of clear and unquestioned cases. For example, Bard et al. show that Chomsky & Rizzi's position that ECP violations are stronger than subjacency violations demands the ability to compare judgments on a reliable quantitative scale. Both agree that the existence of reliable, fine-grained continua in acceptability creates an unresolved problem for a syntactic theory that is essentially discrete.

In his conclusion, Schütze finds himself in general agreement with several working principles for the continued exploration of grammatical judgments proposed in Labov 1975. The first of these recognizes the fact that most of the work of linguistic description will necessarily continue on the basis of elicited judgments.

(7) I. *The consensus principle*:

If there is no reason to think otherwise, assume that the judgments of any native speaker are characteristic of all speakers of the language.

The first principle is subject to modification by the second. The "reason to think otherwise" is disagreement among native speakers, which is certainly not uncommon. In this case, it is suggested that the person who is privy to whatever linguistic theory is developing, the investigator, should not use his or her own evidence.

(8) II. *The experimenter principle*:

If there is any disagreement on introspective judgments, the judgments of those who are familiar with the theoretical issues may not be counted as evidence.

Schütze would strengthen this principle by arguing that the theoretician's own judgments should *never* be used. This seems a bit strong. But certainly this caution must apply when the linguist is engaged in a controversy that depends upon the critical data. Labov 1975 cites a number of such cases where judgments of grammaticality matched theoretical predictions.[1]

In the face of disagreement and uncertainty about grammaticality judgments, most linguists avoid taking the responsibility for using their own judgments to represent the language of the speech community. Instead, they assert that the "facts" submitted pertain only to their dialects, even though it is generally understood that the theory involved governs the language (or all languages) as a whole. It is therefore a crucial question as to whether such idiosyncratic dialects exist: stable systems that govern the speech of individuals in different ways, without any social or geographic correlation.

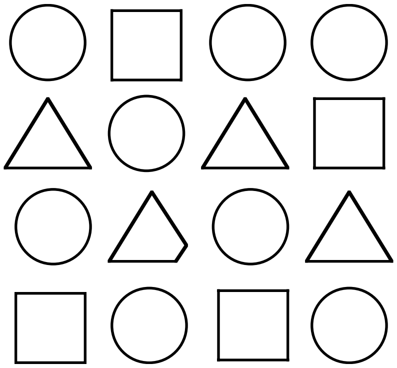
Weinreich, Labov and Herzog 1968 traced the reliance on the idiolect as the fundamental unit of linguistic analysis to Paul's *psychischer* Organismus, a psychologically internalized grammar that generates the utterances of individual speakers. For Paul, the language of the community, or "Language Custom", has no determinate bounds: every grouping into dialects is arbitrary. Thus the language is a composite of stable idiolectal dialects, which may differ widely from each other, and which dialect an individual uses is the product of his or her idiosyncratic mental activity.

A number of such idiosyncratic dialects have been examined, but none of them have been found to be stable or reproducible. The case of the NEG-V/NEG-Q dialects is reported in Labov 1975. Carden (1970) proposed that these dialects were differentiated in the interpretation of such sentences as

(10) All the boys didn't leave

For the NEG-Q dialects (the majority type), this sentence can be interpreted with a negative at the highest node, negating the proposition 'all the boys left', that is, 'most of the boys left, but not all.' For the NEG-V dialects (the minority type), the universal quantifier is the highest node,and the negative modifies the verb, that is, 'for all boys, it is true that if x is a boy, x did not leave'. In the several empirical studies that followed it was found that posing questions of type (3) about sentence (10) resulted in a fairly stable percentage of 70-90% NEG-Q and 10-30% NEG-V (Carden 1970, Heringer 1970, Labov 1972). However, with other techniques, an increasing number of speakers showed a recognition of both interpretations. This led us to the conviction that the idiosyncratic dialects were the products of nonlinguistic factors, that all subjects had the same grammars, including both NEG-Q and NEG-V structures. To demonstrate this, Hindle presented subjects with (9)

(9) If all the squares were triangles, then all the figures would not have four sides. True or false?



This problem demands NEG-V reasoning, since under the NEG-Q interpretation, the hypothesis is irrelevant. Whether or not the squares were changed to triangles, some figures (the circles) would not have four sides. A NEG-Q reasoner would be forced to answer "True". But a NEG-V reasoner would infer that changing the squares to triangles would make it true that no figure had four sides—if it were not for the odd triangle with the clipped corner, which has four sides. The NEG-V reasoner would then answer "False."

Since the question is apparently pointless under a NEG-Q interpretation, and is obviously a test of the ability to notice the clipped triangle under the NEG-V interpretation, 23 out of 24 subjects answered "False."[2] This is one of many experiments that demonstrated that all speakers of English fall have both interpretations available.

The instability of idiosyncratic dialects is coupled with the results of studies in the speech community which lead to a position opposed to that of Paul. The central finding of sociolinguistics is that the community is the stable and systematic unit, and that the behavior of individuals cannot be interpreted without a prior knowledge of the community pattern (Labov 1966). Precise descriptions of the variation found in the community show that it is not idiosyncratic, but rather a statistically predictable pattern that depends upon the social history of the individual. At the same time, all members of large urban communities are found to share a common base for this variation in the mapping of lexical entries onto phonemic classes as well as a uniform pattern of grammatical possibilities (Labov 1989).

Not all grammatical patterns are uniform: there are many cases where intersecting regional and ethnic patterns lead to sharp or gradient differences in both use and responses within the same community. Some of the differences in judgments of grammaticality will turn out to be faithful reflections of differences within the community. It is therefore reasonable to pursue disagreements to see how stable those differences are. Thus the third working principle on grammaticality judgments:

(10) III. *The clear case principle*:

Disputed judgments should be shown to include at least one consistent pattern in the speech community or be abandoned. If differing judgments are said to represent different dialects, enough investigation of each dialect should be carried out to show that each judgment is a clear case in that dialect.

The investigations carried out under this principle require considerable attention to method. The technique used to pursue the consistency of NEG-Q/NEG-V utterances shows that greater reliability can be achieved by having the subject carry out unreflecting semantic interpretations of utterances, rather than perform meta-linguistic tasks. The paradigm of this type of experiment calls for inserting the variable into a larger text, with different variants for different subjects; the story, question or text then continues with sentences that are ambiguous in respect to the interpretation of the variable, until the subject's interpretation is firmly fixed.. The investigator then intervenes with a question that will show the nature of the semantic interpretation that the subject has made, and continues the discussion until that is clear.

Before any such investigation is carried out, one must decide whether the significance of the answer is worth the effort invested. The "Jay-walking" experiment (Labov 1975:122-3) was used to investigate the possibility of a semantic difference between the *get*-passive and the *be*- passive, with and without a purpose clause. A number of studies had shown that the use of the *ge*t passive was increasing steadily among younger speakers, so that it became a important to know whether this was a case of semantic change or purely a formal shift of auxiliary. But in many cases, the acceptability of a given sentence may be only one of hundreds of pieces of evidence used in the formal linguistic argument, and there may be little profit in deciding if the judgment of the theoretician were right and wrong.

One example of a critical case of acceptability that has played an important role in syntactic theory over the years involves the contraction of *want to* in sentences such as

(11) Who do they want [PRO to visit [$_{NP}$ e]]
-> Who do they wanna visit?

In Chomsky 1981, one of the characteristic properties of the PRO which is the subject of *visit* is that it is invisible to the PF-component that contracts *want* and *to* to *wanna*. PRO is relevant only to the LF-component, or to morphology [p. 23]. However, the trace e is said not to be so invisible, but blocks contraction in the same way that a noun phrase would. Thus contraction is said to be possible in (11) but not in (12)

(12) Who do they want [[NP e] to visit Paris]
* -> Who do they wanna visit Paris?

The accepted fact that the contraction is ungrammatical when a subject has been extracted from between *want* and *to* has been a consistent theme of syntactic discussion since 1981. If this observation should not be correct, then considerable changes would have to be made in the conception of the trace e, or of the relations of the PF-component to the LF-component. Since I shared with many others the reaction that such contractions are possible in spontaneous speech, I introduced the problem of developing an experiment using unreflecting interpretation on this point to a class on experimental methods in linguistics. Karins and Nagy (to appear) developed such an experiment and applied it to sizeable groups of subjects. The spoken text that they played to subjects read as follows:

(13) A mailman is walking down the street. He sees a big roll of money drop from an old lady's bag in front of him. The old lady doesn't realize what has happened. The mailman goes for the money. The old lady sees this, and goes for the money as well. They start struggling with each other. The old lady yells for help.
A policewoman and a big strong guy are standing nearby.
(i) Which one would you **want to** help?
or
(ii) Which one would you **wanna** help?

This text follows the general format of presenting an ambiguous situation where *which one* might refer to a choice between (a) the mailman and (b) the old lady (potential objects of help) or to the choice between (c) the policewoman and (d) the big strong guy (potential subjects of help). In answering the question, listeners focus eiher upon the choice between (a) and (b), or on the choice between (c) and (d), rather than on the decision to consider (a) vs. (b) on the one hand, or (c) vs. (d) on the other. The semantic interpretation of *want to* or *wanna* which focuses their attention on one of the two pairs is made unconsciously and automatically. In their responses, subjects reveal whether they interpreted the WH- form as co-indexed with a moved object or subject of *help*.

The two recorded texts were played to three different undergraduate classes at the University of Pennsylvania. Two small classes (N = 25) were presented with the text using sentence (i), the uncontracted form of the question. The third class (N = 57) was presented with the text using sentence (ii), the contracted form. Additional data, from speakers across the country, was collected by students of a class on American dialects as part of a larger questionnaire investigating variation in American English (N = 34, 14 uncontracted, 20 contracted). Karins and Nagy's overall results are shown in Table 1.

Table 1. Responses of all speakers to the *wanna* experiment
[from Karins and Nagy 1996)

Question	Object interpretation	Subject interpretation	Total No.
(i) want to	7 (18%)	32 (82%)	39
(ii) wanna	50 (65%)	27 (35%)	77
TOTAL	57(49%)	59 (51%)	116

These results appear to give strong confirmation to the general theory that differentiates PRO from the trace.

The construction of the text naturally favors the subject interpretation. The sentence immediately preceding the question refers to those who might help, and there is a strong tendency for pronouns like *one* to refer to the nearest antecedent. The results for the 39 subjects who heard *want to*, which we assume does not favor subject or object interpretation, are consistent with this expectation: they favored the subject interpretation by more than 4 to 1. This textual bias was strongly reversed for the 77 subjects who heard contracted *wanna*: they favored the object interpretation 2 to 1.[3] Therefore the constraint against the subject interpretation with contracted *wanna* was powerful enough to overcome the textual bias in favor of that interpretation.

This confirmation of the theoretical principle is the type of confirmation obtained in the investigations reported in Schütze 1996 and in Bard et al 1996, and might lead us to the conclusion that there is no fundamental problem with the intuitive judgments of grammaticality in Chomsky 1981 and echoed by many others since. There is no doubt that the *wanna* constraint on subject interpretation is a real one.

On the other hand, we have to face the fact that 27 subjects, one third of those who heard sentence form (ii), acted in a contrary manner. We could say that their interpretations were performance errors; or that they decided unconsciously to ignore the ungrammaticality of the subject interpretation in the light of the discourse bias of the text. A more likely possibility is that we are dealing with a linguistic variable, and the *wanna* constraint is itself variable, a gradient phenomenon. Whether or not the theory that states that the empty subject category is equivalent to a full noun phrase in the PF-component could be modified to deal with gradient behavior is an open—and difficult—question.[4]

Principle III deals with disagreement among judges on the meta-questions (1-5). There is also disagreement between the results of observation, gathered by inquiries under question (6), and the results of inquiries using questions (1-5). In the course of their inquiry, Karins and Nagy made observations of language in use, and gathered the following examples of speech production using *wanna* contraction with subject deletion:

(14) You'll have to decide who you wanna lead this country. (President George Bush, 10/13/92)
(15) These are the people I wanna be there. (observed by Eugene Buckley)
(16) Who do you wanna represent you? (question asked of Anthony Kroch in class, and responded to immediately)
(17) What if you have a couple of columns you don't wanna be cut? (C. Cieri, 2/23/93)
(18) That's the type of mother you wanna visit more often. (S. Boas, 9/19/93, referring to the mother visiting)
(19) They'll pick the candidate they wanna win. (S. Strassel, 6/22/95).

To preserve the discrete character of the theory involved, it would be natural to assign (14-19) to the status of performance errors. On the other hand, they may be taken to confirm the view that the variable results of the Karins and Nagy experiment reflect a variable constraint that operates in the production of sentences as well as in perception. These observations alone do not give us a decisive basis for resolving this issue; they are in the authors' terms, "anecdotal." They were not gathered under the principle of accountability that would tell us how many times subject deletion was performed without *wanna* contraction, so we cannot compare the experimental results with the results of observation. There are available, however, a number of studies that allow us to compare judgments of grammaticality with the pattern of production in a systematic fashion, and the rest of this paper will deal with reports of this kind. To interpret these results, some general principle is needed to guide us in comparing subjective and objective data. The principle proposed in Labov 1975, which I will follow here, is:

(20) IV. *The principle of validity*:

> *When the use of language is shown to be more consistent than introspective judgments, a valid description of the language will agree with that use rather than with intuitions.*

Many linguists, perhaps the majority, may disagree with me here. This is a critical issue that goes beyond methodology: it rests upon our conception of what language actually is. This principle of validity differentiates the materialist approach to language that is at the heart of the sociolinguistic perspective, and the idealist position that is adopted by many formal grammarians. I believe however that the data to follow will be justify (20) as a useful working principle

The results of Bard et al 1966 and Karins and Nagy 1966 generally conform to the expectations projected by the linguistic theories being tested. Let us suppose that with proper precautions we can obtain an array of judgments of acceptability from the general population that confirms statistically the import of most theoretically informed judgments. That would not, as I see it, provide a solid empirical base for linguistics. The essential problem is that in another set of cases, judgments of naive subjects are massively skewed from actual use, reversing the linguistic facts. It is not the number of these cases that poses the problem, but the fact that we cannot predict when and where they will occur. We are in the position of using a ruler that displays random error. We would like to say that structure X measures 5 points higher on a scale of grammaticality than structure Y, with a probable error of ± .5. Instead, we have to say that there is a distinct but unknown probability that the 5 on the scale is actually a 2, or a -8. To the extent that such random error in the elicitation of introspective judgments persists in our data, we are building on insecure foundations..

The rest of this paper represents an effort to locate these errors. At a number of points so far, I have referred to "introspections" rather than "intuitions," following the general understanding that we have no guarantee that our subjective reactions are in fact the intuitions that govern language. But since "intuitions" is the more common term, I will use this term more generally in what follows, with the understanding that the term refers to the elicitation of introspective judgments.

The mismatch of intuition and behavior in morphology: the case of positive anymore.

The current Telsur project at the Linguistics Laboratory at Pennsylvania is devoted to mapping linguistic changes throughout the U.S and Canada, for the *Phonological Atlas of North America.* In these telephone interviews we include a number of grammatical variables, such as the alternation of *The car needs washing* with *The car needs washed,* or the use of positive *anymore* in sentences like *Cars are sure expensive anymore..*

We would like to extend these grammatical inquiries into a much wider range of regional phenomena. However, results so far confirm our earlier experience that responses to direct inquiries on grammatical patterns are of questionable reliability. The erratic scattering of responses obtained through introspection has little relation to the clear regional patterns produced by mapping behavior, whenever it is frequent enough to be registered in an hour's conversation on the telephone.

In Philadelphia, we have had ample opportunity to study responses to questions about positive *anymore* (Labov 1972, Hindle 1974). This variable represents the use of *anymore* in positive sentences to mean that the situation described was not true at some time in the past and is now true: roughly equivalent to 'nowadays'. Since 1972 we have collected many hundreds of examples of the use of *anymore* in spontaneous conversation. It is used freely in all sections of the Philadelphia white community, from lower class to upper class; in all these instances, and in our many observations throughout the Midland area, we find no evidence of social stigma.

Yet introspective responses to questions about *anymore* are very erratic indeed. In 1973-4, we identified 12 speakers who used positive *anymore* freely though responses to questions of types (1-5) were entirely negative. Jack Greenberg, a 58-year-old builder raised in West Philadelphia, gave introspective reactions that were so convincing that I felt that I had to accept them as valid descriptions of his grammar. Yet two weeks later, he was overheard to say to a plumber,"Do you know what's a lousy show anymore? Johnny Carson."[5] A 42-year-old Irish woman said, "I've never heard the expression." Earlier in the interview she had said, "Anymore, I hate to go in town anymore," and a short time later, "Well, anymore, I don't think there is any proper way 'cause there's so many dialects."[6]

Whenever we have been able to maintain continued contact with a white Philadelphia speaker, we find that he or she uses positive *anymore* sentences when a favorable contact arises. From this and other evidence, we came to the conclusion that positive *anymore* was an invariant property of the grammar of the Philadelphia speech community. As our techniques for eliciting judgments improved, we obtained a higher and higher percentage of consistent responses.[7]

The problem of intuitive responses to positive *anymore* is acute: we do not know why speakers find it so difficult to recognize their native grammatical patterns.

The mismatch of intuition and behavior in tense and aspect: the case of Black English BIN

Some of the most striking examples of the failure of intuitions to give a reliable view of linguistic structure are found in the study of African American Vernacular English [AAVE], where we have more extensive records of spontaneous speech than for any other dialect. Much of this work has focused on the systematic study of speech production in syntax, morphology and phonology. However, this is very difficult to do for AAVE aspect particles : invariant *be, done,* stressed *been, be done,* and *been done.* These are the critical features that separate the grammar of AAVE most sharply from other dialects. On the positive side, they carry complexes of semantic features that are not found in any other dialect (Rickford, Dayton 1996, Labov 1996). On the negative side, they do not carry deictic indications of tense, and can be used with any time reference, and do not show any of the syntactic properties associated with the presence of raising a verb or auxiliary to INFL: inversion, tag formation, adverb placement, affixal negation. They are difficult to study systematically in production because

(a) they are concentrated in interactive use in vernacular situations. Except for invariant *be,* they occur rarely in interviews and are not observed at all in public displays of black speech on the mass media or in literature.

(b) they appear to be optional features of discourse: there is no agreed analysis of closed sets that would permit us to observe when these aspect particles do not occur as well as when they do occur.

In this section, will concentrate on stressed béen, usually referred to as BIN, using two data sets. This is a pre-verbal particle that is uttered with pronounced stress, usually with a low tone. The major source of information on the use of BIN in production is Dayton 1996, a report of four years of participant observation in an all-black community of West Philadelphia. Dayton's data base of 490 uses of BIN was recorded by her in writing, with the full contextual details of the situation involved. The total number of observations is an order of magnitude greater than all other reports of the use of BIN combined. Sentences like (21-29) were written down by Dayton shortly after they were spoken.

(21) Woman A, 20's: Can we have hot dog rolls?
Woman B, 30's: If you all want hot dog rolls, you shoulda BIN took them out (of the freezer). It's too late for hot dog rolls. Get the bread!
(22) Man,, 20's (Playing Uno): If I was playin' Tyrone, I woulda BIN beat.
(23) Woman, 30's: I BIN broke life and death down. 'I figured out the meaning of life a long time ago.'
(24) Woman A, 30's: Do you like the new miniskirts they got out?
Woman B, 30's: Miniskirts Bin out! Miniskirts BIN out! Miniskirts has BIN out!
(25) Woman A, 30's: What are you doing? mailing that to Calvin?
Woman B, 30's: No, I BIN mailed that. I mailed it last week.
(26) Woman A, 30;s: When she hit him? Yesterday?
Woman B, 30's: *Uh, uh.* She BIN hit him about three weeks ago.
(27) Woman, A, 30's: When you get this coat?
Woman B, 30's: I BIN had it.
Woman A, 30's: When you get it?
Woman B, 30's: I BIN had it; I Bin had it; I BIN had it since a month ago!
(28) Woman A, 20's (Watching TV): See, they dead.
Woman B, 20's: They BIN dead.
Woman A, 20's: No, they didn't; they just died.
I seen that one (pen) Tyrone gave you. I BIN had this one before he gave that to you.
Man, 20's: Three months ain't such a long time.

These examples illustrate the general meaning of BIN: 'anterior, remote, perfective'. It is used to refer to events that occurred (psychologically) a long time ago;[8] this remoteness is explicitly asserted in (25), (26), and (27). It is acknowledged and denied in (28) and (29).

The perfective sense of present relevance, which BIN shares with the general present perfect, amounts to a claim that the state of affairs referred to is still true (with stative predicates as in (23,24,27,28) or that its effects continue to prevail (with non-stative predicates as in (21,22,25,26,29). However, BIN does not share the present perfect sense for stative predicates that implies (without temporal modification) that the condition no longer holds. Thus (30) and (31) contrast in this sense: (30) implies that she is longer married, (31) that she is still married.

(30) She (ha)s been married.
(31) She BIN married.

Rickford 1975 used this contrast in an investigation of judgments on BIN by black and white subjects in West Philadelphia (the same general community as that studied by Dayton). A series of questions of type (1-5) begain with (32).

(32) Someone asked, "Is she married?" and someone else answered, "She BIN married"
Do you get the idea that she is married now? Yes___ No ___.

The questionnaire was administered to a broad range of adults in the West Philadelphia outside of the university. In general, they gave the Remote Perfective interpretation in 85% of the cases, and while the white subjects he interviewed did so only 37% of the time.

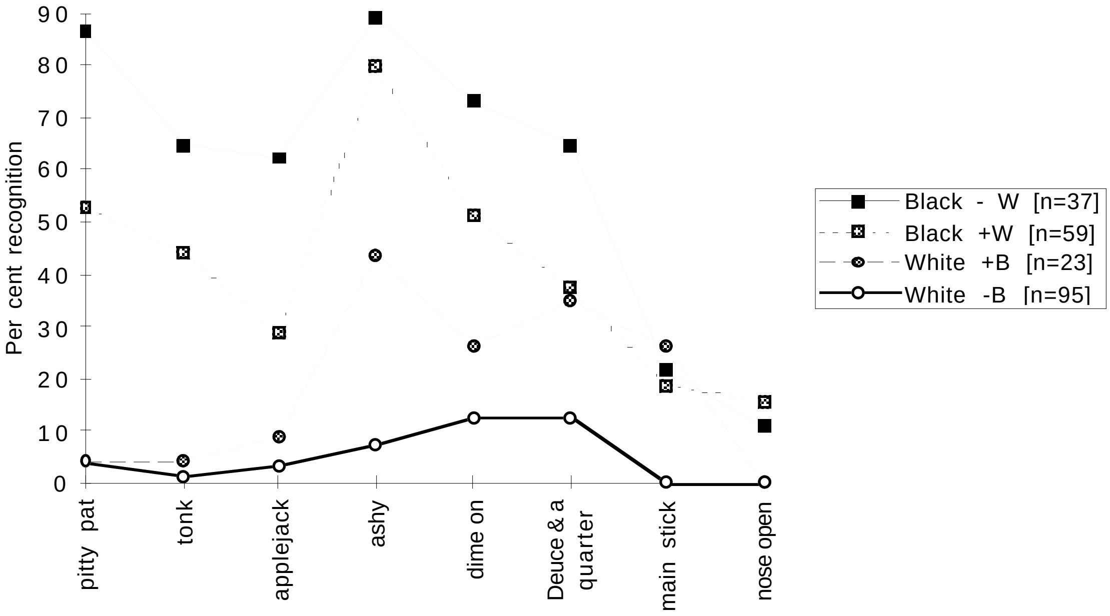
Rickford also obtained an order of acceptability for fourteen syntactic contexts of BIN. Dayton 1996 compares this ordering of acceptability with her own actual observations of frequencies of use in Table 2.

Table 2 Acceptability ratings and corresponding number of tokens for BIN sentences. [from Dayton 1996]

Acceptability Rating (Rickford 1975) (1=most, 14=least)	BIN Sentence (Dayton 1996)	No. of Tokens
1. ___ Ved	They BIN ended that war.	302
2. ___ Ving	I BIN knowing him.	52
3. ___ NP	He BIN the leader.	0
4. ___ 've had	I've BIN had that car.	0
5. ___ Pass.	The chicken BIN ate.	0
6. ___ knew	I BIN knew your name.	23
7. ___ got Pass.	He BIN got messed up.	1
8. ___ have	I BIN have that.	0
9. ___ Adj	She BIN nice.	32
10. __ Modal	I BIN could do that.	0
11. __ done	He BIN done gone.	55
12. done __	He done BIN locked up.	5
13. ___ bin	He BIN bin gone.	1
14. have __ had	I have BIN had that.	0
Total		471

In general, use and acceptability judgments are not badly matched, though not precisely enough so that we can use one to correlate the other. The most common context, BIN with a past tense verb, was also rated the most acceptable, and the third most com-

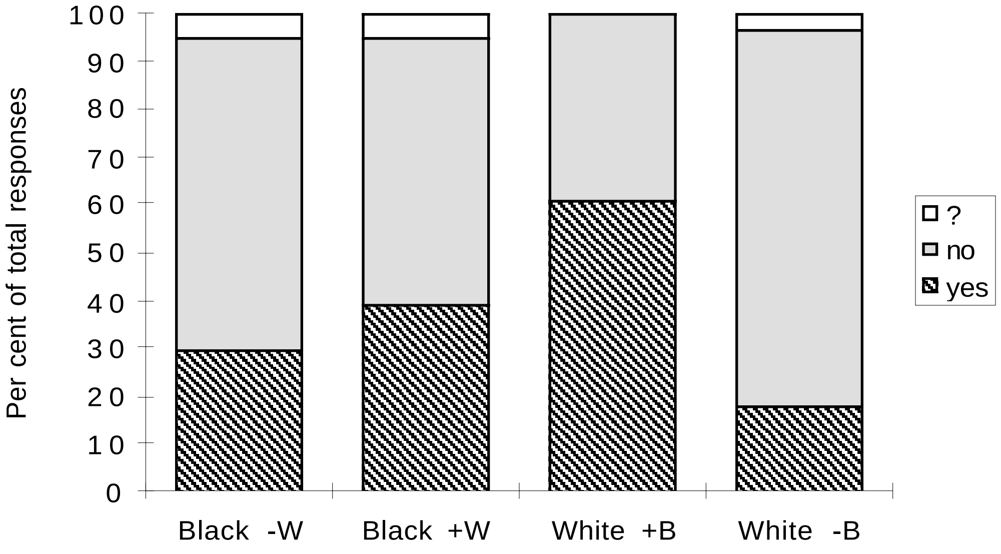
Figure 1. Recognition of AAVE lexical items in the 1984 student survey



Perception of the meaning of BIN.

To study subjective responses to BIN, we included in this study the key first question in Rickford's study (32.1). It was expected that those unfamiliar with Remote Present Perfect BIN derive "She been married" from general English "She has been married", but speakers of AAVE would recognize the stressed low tone on BEEN and assign the meaning 'still married'.

Figure 2. Does "She BEEN married" imply that "She is married now'? [1984]



The results for this question, shown in Figure 2, are quite startling. For the black subjects, the level of “yes” response to the question is much lower than that obtained by Rickford, only slightly higher than the white population with no black contacts. On theother hand, the group of 23 whites with extensive black contacts shows a level of "yes"response significantly higher than both groups of blacks ($\chi^2 = 4.2$, $p < .05$). Since there is no reason to believe that the whites have actually incorporated stressed BIN into their own grammars — no actual use of stressed BIN for this group has been reported — we can only interpret this as a report of how they have learned to interpret the speech used by the blacks they have been associated with. It is therefore not self-report, but rather a report of observations. This is only one of many examples where observation from the outside can be more accurate than intuition.

How can we account for the low percentage of “yes” reports by blacks who come to the university from black environments? It seems likely that their grammatical intuitions have been inhibited or overlaid by their current strong contacts with the general American pattern. This pattern would not be expected to inhibit their knowledge of *tonk*and *pitty*-pq5, but it apparently acted to repress their semantic interpretations of stressed BIN.[9]

Mismatch of intuition and behavior in negation: The case of AIN’T in the preterit

This inhibition of the specific features of AAVE is not confined to stressed BIN. Let us consider the use of *ain't*, which separates black and white communities with equal clarity in production. While all but the standard variety of English uses *ain’t* for the negative present auxiliary in place of *isn't*, *amn't* or *aren't*, and for the negative of the present perfect in place of *hasn't* and *haven't*, only AAVE uses *ain't* in the preterit, in place of *didn't*. To the best of my knowledge, no white speaker has been heard using *ain't* in sentences like (34-36) from the South Harlem study.

(34) I ain't see the fight, and I ain't hear the fight. [13, Jets, NYC].
(35) ...so I told him I ain't pull it. [15, Lame, NYC].
(36) Well, he didn't do nothin' much, and I ain't neither.
[12, TBirds, NYC].

Table 3 shows the alternation of *didn't* and *haven't* in South Harlem for a number of different groups, indicating that the vernacular use centers about 50%.[10]

TABLE 3
COMPARATIVE USE OF AIN'T AND DIDN'T IN THE NEGATIVE PAST
FOR AAVE PEER GROUPS AND OTHERS IN SOUTH HARLEM

	Style	*ain't*	*didn't*	*% ain't*
T-Birds (12)	A	8	17	32
	B	11	23	32
Cobras (16)	A	38	44	46
	B	8	10	45
Jets (32)	A	29	42	41
	B	60	64	48
Oscar Bros (4)	A	20	20	50
	B	19	26	40
Lames (10)	B	10	33	23
Inwood (white) (8)	A	1	22	04

B	0	37	00

-- Labov, Cohen, Robins & Lewis 1968

Now let us consider how the 217 subjects of the 1984 survey reacted to questions about the grammaticality of *ain't*. The second question on the survey is shown in (37), with the original random order of the items re-arranged for the purposes of analysis.

(37) Some people use *ain't* in every-day language. Can you or the people you've heard use *ain't* in:
(General non-standard)
___He ain't too smart. [Present copula]
___He ain't been doin' that too long. [Present perfect]
___He ain't gonna go there tomorrow. [Present copula in periphrastic future]
(General Southern)
___Ain't nobody know my name. [Copula. deleted dummy IT]
(AAVE only)
___He ain't see her yesterday [Preterit]
___Why ain't he do that? [Preterit]
(Ungrammatical)
___I ain't really wanna do that. [Present]
___She ain't know that now. [Present]
___ You better ain't do that. [Non-finite]

The first three items are general non-standard uses of *ain't*. The fourth item is structurally ambiguous. It can be derived from (38) with deletion of dummy *it* (Northern *there*). This is general non-standard Southern States English. The other possibility is negative inversion of (39), which is grammatical only in AAVE with a preterit reading. The last three items are ungrammatical in any system.

(38) It ain't nobody know my name
(39) Nobody ain't know my name

Figure 3 shows the percent acceptance of *ain't* in these nine contexts. For the general non-standard items, there is a fairly uniform situation, with a high degree of recognition by all but the whites with little black contact. For the ambiguous item, *ain't nobody know*, there is an intermediate situation. But there is no clear differentiation between responses to the AAVE *ain't* for *didn't* and responses to the first two ungrammatical items. Only the last item, the use of ain't for non-finite not, receives a universal rejection.

Again we observe that whites with extensive black contacts show more recognition of the vernacular pattern than any other group for the general non-standard *ain't* and two of the AAVE forms. It is only for the third AAVE item, *ain't see her* yesterday that this group fails to show greater recognition than the balck subjects. On the other hand, the whtes with no black contacts are clear on the negative side: they are less confused about these and other non-standard uses, part of their general lack of recognition of *ain't*.. On the other hand, there is no differentiation of blacks by the amount of white contact.

The last question in the survey was (40), which sought to elicit differences between black and white subjects by another route.

(40) How many meanings can you find for *She ain't like me?*

The results were quite striking and unexpect, as shown in Figure 4. The groups are not particularly different from one another: the between-items variation is much greater than the between-group variation. Almost everyone gave 'isn't like' as a meaning, and almost no one gave 'didn't like'. (Again, only a few of the whites with strong black contacts distinguished themselves in this respect). What was more surprising is that about 50% of the

Figure 3. Judgments of grammaticality of ain't in nine contexts in the 1984 student survey

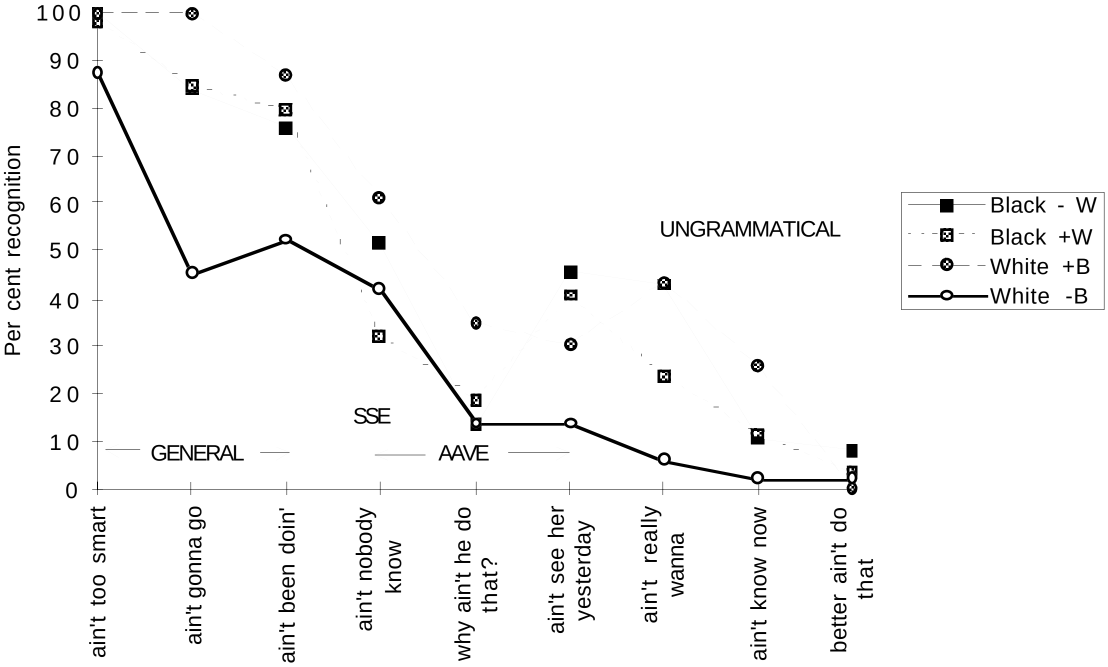
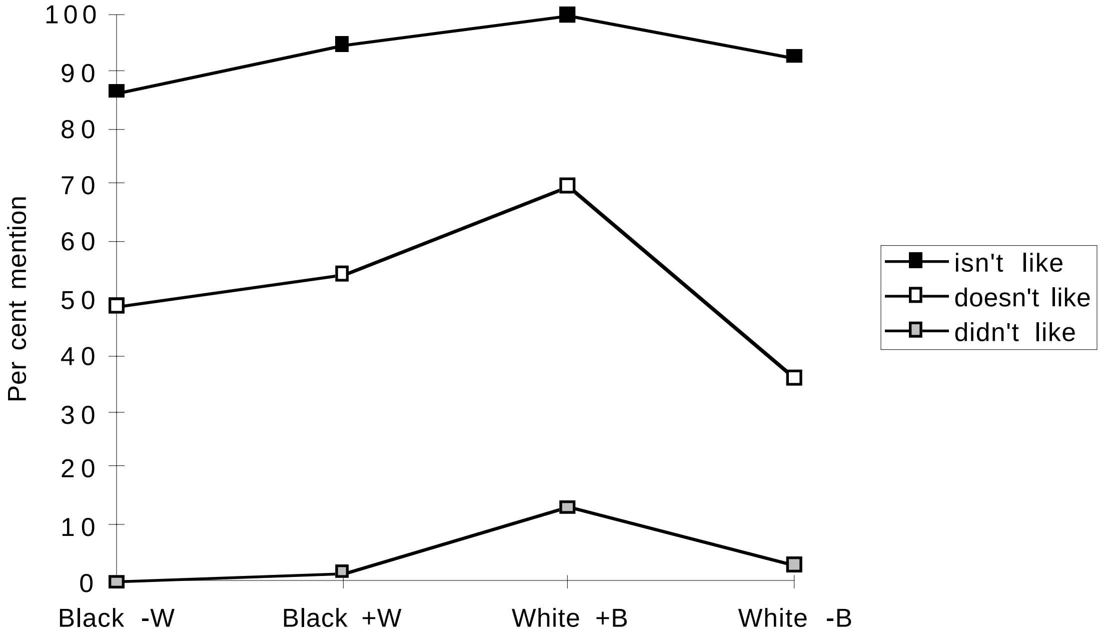


Figure 4. Multiple interpretations of "She ain't like me" by four racial groups [1984]



mon, BIN with the progressive, was rated second. But the 3rd, 4th and 5th most highly rated did not occur at all, and the second most common was rated 11th in acceptability. Because we cannot close the envelope of variation in Dayton's data, massive though it is, by saying how many times BIN did NOT appear in a given environment, we cannot say how significant the matches and mismatches between judgments and production are. It is possible (though not likely) that the contexts for constructions 3, 4, and 5 simply did not arise, and this would explain the 0's in the Dayton column. It would be more difficult to explain the low rating given to *BIN done gone*. On the other hand, Rickford's and Dayton's observational data leave not the slightest doubt that stressed BIN is an integral part of African American Vernacular English.

The student survey of BIN

Let us now consider a study of grammatical judgments that is more typical of the data bases that are available to most linguists: a survey of grammatical judgments carried out by students in an introductory class a few years after Rickford's study. The 214 subjects of the survey were mostly students at Penn, though they included some outside of the university community. The subjects can be divided into four groups, based on the amount of cross-ethnic experience they had, as reflected in the answers to two questions on the survey: percentage of the other ethnic group in high school, and number of friends of the other ethnic group.

(33) Black -W [37]
High school with no more than 30% whites and reporting no more than 5 white friends.
Black +W [59]
High School with more than 30% whites or reporting more than ten white friends.
White +B [23]
High school with 30% black or reporting more than 10 black friends
White -B [95]
High school with less than 30% black.

The lexical differentiation of the population

To get some idea of how intimately these four groups are linked to African American Vernacular culture let us examine first the results for lexical items on the questionnaire in Figure 1.

The first three terms on the left are specific to the black community and divide the subjects into two groups by race: *tonk* and *pitty-pat* are card games universal in the AAVE community, but almost unknown to whites. The hat known as an *applejack* falls in the same class. Blacks with minimal white contacts show uniform recognition, while blacks with more white contacts are intermediate.

A second group of terms show evidence of passing over from the black to the white community, and divide the subjects into three or four groups depending on their degree of contact: *ashy* (appearance of dry skin in the winter'), to *dime on* ('inform'), and *Deuce and a quarter* ('Buick Electra 225').

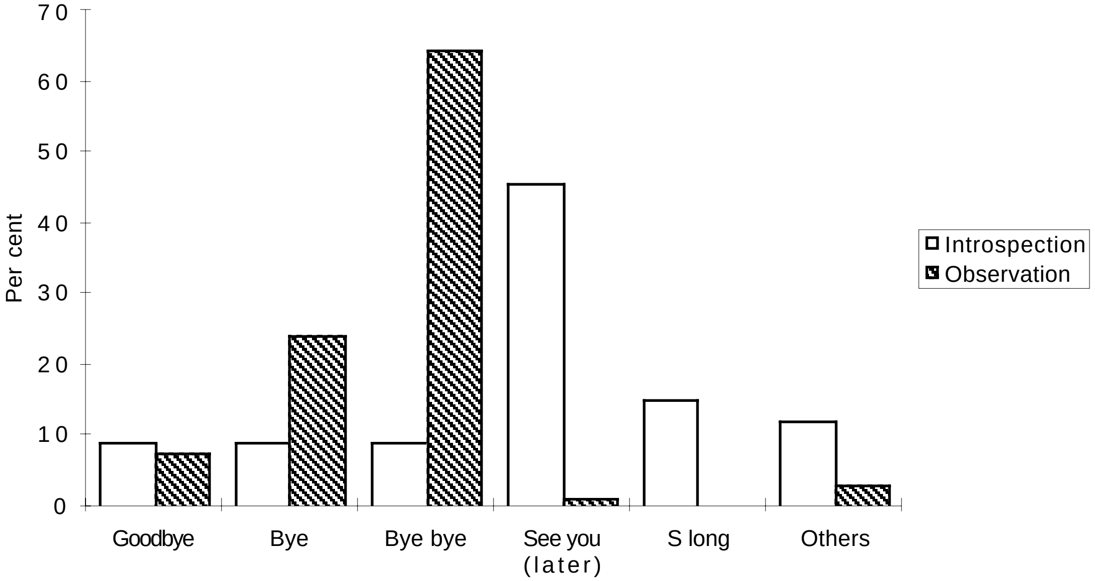
Main stick appears to be passing out of use, and is recognized by only 20% of blacks and whites with black contacts. *To have your nose open* ('to be deeply in love') is also passing out, and is altogether unknown to white subjects.

subjects gave 'doesn't like,' which is never actually heard. We must conclude that a question of this type, relying on the detection of ambiguity, is less likely to give realistic responses than the other types mentioned so far.

The mismatch of Intuition and behavior in discourse: the case of bye bye.

There appear to be many more mismatches of observation and intuition in discourse than in morphology or syntax. One such case that I have investigated concerns the last utterance in a conversation—in person or on the telephone. People are generally unaware that their most common form of final leave-taking is *bye-bye*. When this is demonstrated to them, they usually react with unbelief, since *bye-bye* is associated with childish or effeminate behavior. Figure 5 shows the contrast between the intuitions of 33 subjects and a set of observations made at the same periods that the questions were asked.[11]

Figure 5. Introspections by 33 subjects of their most common last utterances compared to observations of 234 last utterances in Oct 1970 and Nov 1971



The instability of American last utterances can be ascribed to the general patterning of ritual behavior in our culture. In general, leave-taking is a difficult and problematic social event. It is important that each participant in the interaction assure the other that they remain in a right relationship, that nothing has transpired in the last interaction to change that status. This assurance is given in American culture by a demonstration of sincere and warm regard for the other. The complication results from the fact that at the same time, ritual behavior is associated with a lack of sincerity; sincerity is demonstrated by spontaneous behavior, which contrasts sharply with ritual.

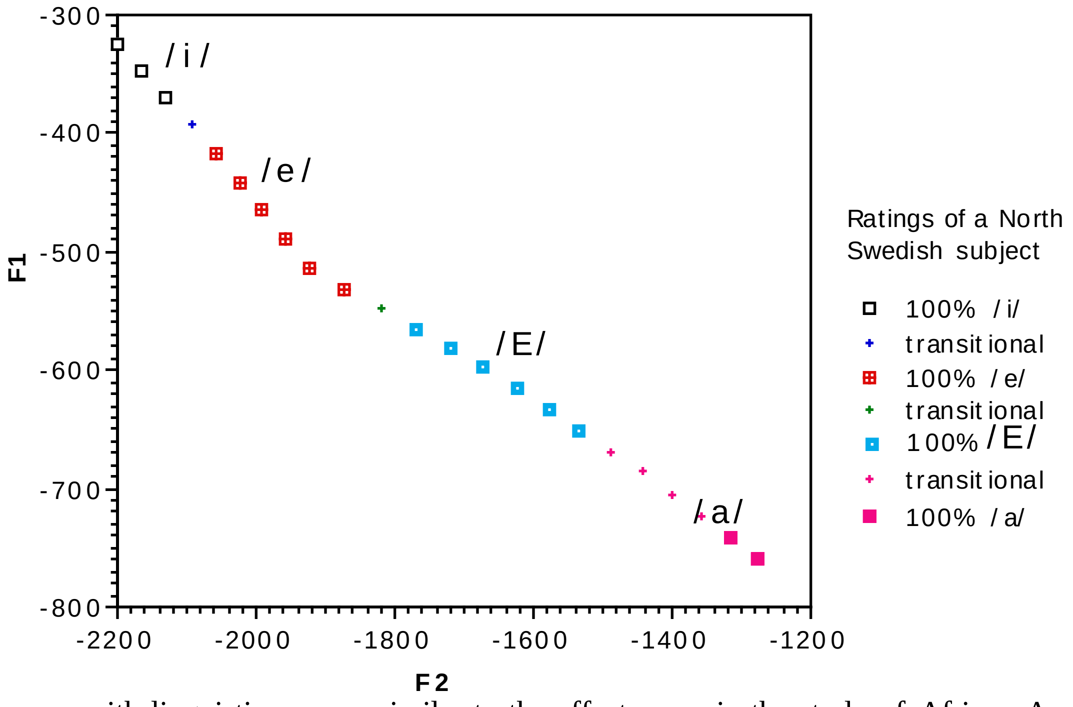
The result of this cultural configuration is that as soon as a form of leave-taking rises in the level of social awareness so that it is recognizable and reportable as a ritual, it is no longer a suitable means of leave-taking. It is therefore replaced by other forms that are accepted as spontaneous because they are not recognizably ritual. The gradual reduction of *goodbye* to *g'bye* to *bye* is one form of response to this dialectic. Another is the replacement of *bye* with the hypocoristic *bye-bye*, which also offers a variety of phonetic forms to signal variable degrees of warm regard. In recent years, the warmest of these, with a lax high back rounded vowel, has been steadily gaining ground.

Mismatch of intuition and behavior in phonology: the case of the Swedish front vowels.

In the domain of phonology, the asymmetry of production and perception has been a major topic of inquiry since 1972, when the first case of near-mergers was discovered in the vowel system of New York City. Though naive speakers and linguists alike heard r-less *source* as identical to *sauce*, formant measurements showed a consistent and statistically reliable difference between the two word classes—in the same orientation that r-pronouncing dialects showed (Labov, Yaeger & Steiner 1972, Ch. 6). In this and other cases of near-merger, native speakers consistently make a distinction that they can not label in minimal pair or commutation tests.

In this area of investigation, we can come closer to identifying the causes of the mismatch between perception and production, and so have a better chance of predicting when intuitions will fail. Janson and Schulman (1983) encountered such a case in their study of the categorization of Swedish vowels. Figure 6 shows the series of 20 synthetic stimuli that they submitted to Swedish speakers for identification. Most Speakers of Northern dialects maintain a clear distinction between four short front vowels /i,e,E,a/ while Stockholm speakers merge the second and third. The symbols identified in Figure 6 show the tvowels that a typical Northern speaker identifies 100% of the time as /e, e, E/ or /a/. The stimuli that produced variable behavior are indicated as small dots.

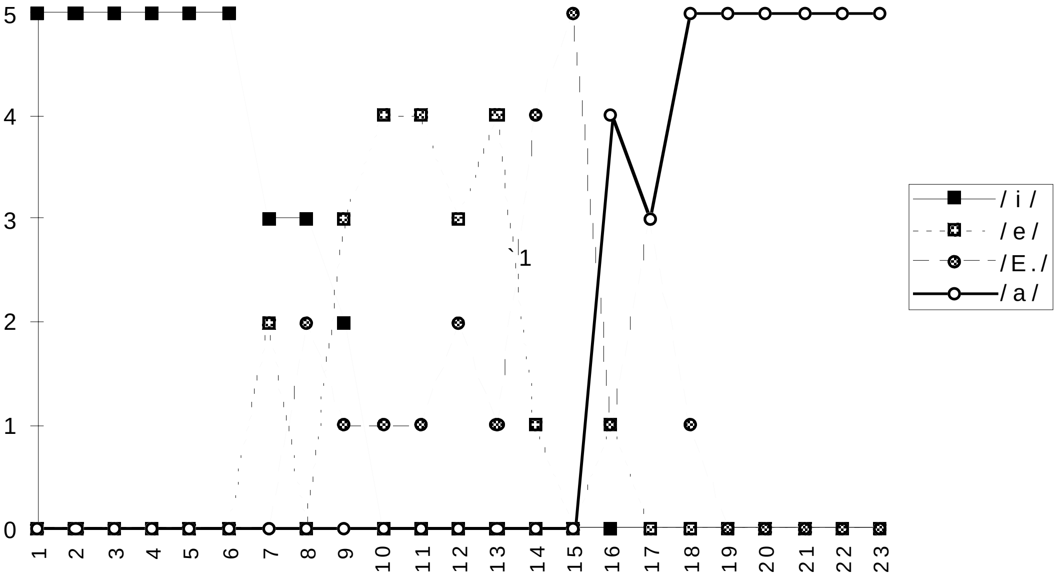
Figure 6. Synthetic stimuli submitted to Swedish speakers, with areas of 100% ratings of a North Swedish subject [Janson & Schulman 1983].



norms with linguistic norms, similar to the effects seen in the study of African American Vernacular English.Janson and Schulman note that in another experiment, they identified the 23 stimuli as syllables of American English. The result was a categorization much closer to that of other Northern subjects. The region of uncertainty was reduced from an average of 7.1 vowels to 3.8.

When the same stimuli are used with speakers of the northern Lycksele dialect, which also has four phonemes, results such as Figure 7 regularly appear. It is immediately apparent that for this speaker, /e/ and /E/ are the *same* in some subjective sense, even though in his own speech, they are produced as distinct entities. The pattern of Figure 7 resembles the pattern of categorization of a Stockholm speaker, who has only three front short vowels, and it seems evident that the Lycksele speaker has substituted the Stockholm categorization for his own in his responses to the experiment. Since the Stockholm dialect is socially dominant, this substitution can be regarded as a case of the interference of social

Figure 7. Categorization of 23 synthetic vowels by a speaker of the Lycksele dialect [Janson & Shuman 1983]



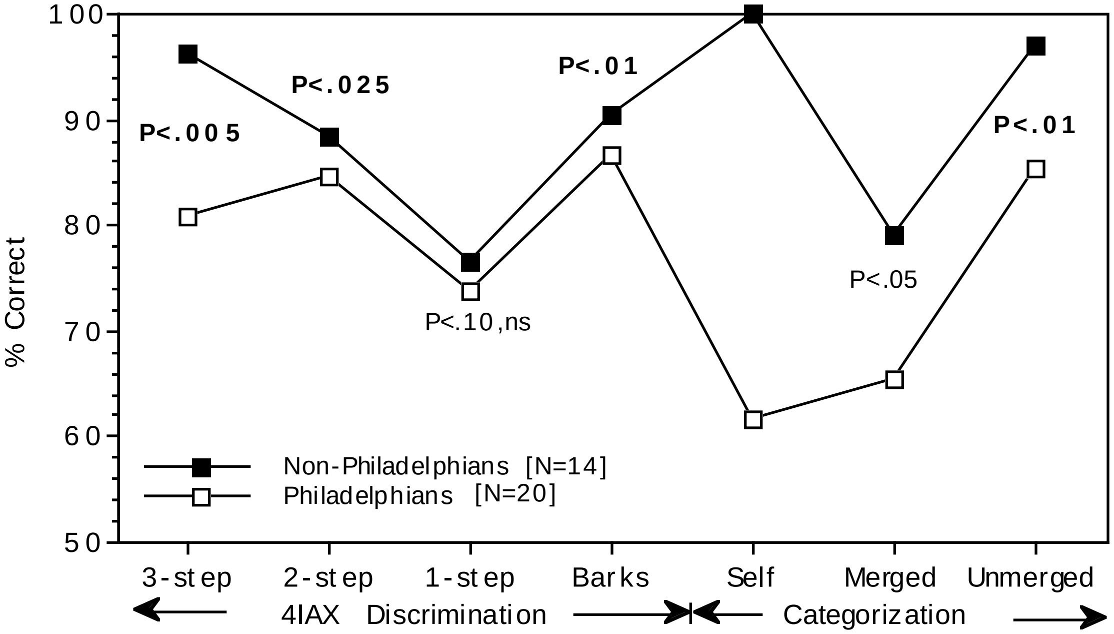
Mismatch of intuition and behavior in phonology: the case of *ferry* and *furry*.

Labov 1994, Ch. 12 presents a number of other cases of near-merger, and a great many other have been identified in the literature since 1972 (Labov 1994, Di Paolo 1987, Di Paolo and Faber 1990, Faber and Di Paolo 1995). The one that we have studied most closely is the Philadelphia near-merger of the /e/ and /U/ before intervocalic /r/ in *ferry* and *furry, merry* and *Murray,* etc (Labov 1994: Ch. 14).

There is a great deal of individual variation within the Philadelphia speech community: about one third of Philadelphians make a distinction between *ferry* and *furry,* a third have a complete merger, and a third show a near-merger, where they maintain an F2 distinction of about 200 Hz, with small overlap, but label the two categories as the same, in their own speech or the speech of others. Figure 8 shows the results of a series of categorization and discrimination experiments with Philadelphians and non-Philadelphians designed to answer the question: where do the psycho-acoustic abilities of Philadelphians differ from those of non-Philadelphians in their perception of this contrast?

The three categories on the right show the results of commutation tests, where subjects are asked to identify a random series of words which were read by a speaker as *ferry* or *furry*. In the SELF commutation test, each person hears his or her own reading of the randomized word list, starting at some point unknown to them. Here all 14 non-Philadelphians show 100% success, and Philadelphians are close to chance. But we do not know whether this is due to the pronunciation or the perception of the Philadelphians. In the two following commutation tests, all subjects hear the same stimuli: read by a Philadelphian with a near merger ('MERGED') and read by a Philadelphian with a clear distinction

Figure 9. Mean results of discrimination and categorization for Philadelphians and non-Philadelphians



('UNMERGED'). Here the mean difference between Philadelphians and non-Philadelphians is less dramatic. These two tests indicate that about 40% of the difference between the two groups in the SELF commutation test is due to a difference in the ability to label the differences.

The four items on the left are experiments designed to bypass the labelling behavior that minimal pair and commutation tests evoke, and detect absolute differences in psycho-acoustic ability. The experiment is a 4IAX discrimination tests, where subjects hear twopairs of words, and are asked to say which of the two pairs shows a difference. The stimuli are all derived from the same natural production of Murray, where the height of the second formant is systematically varied so that the different pair varies by steps of 100 Hz. The 3-step experiment at extreme right, the 'different' pair differs by 300 Hz; in the 2-step experiment by 200 Hz, and in the one-step by 100 Hz. The fourth discrimination experiment uses only the vocalic portion of the word; it is considerably shorter, and sounds like a non-linguistic bark.[12]

The 3-step, 2-step, 1-step pattern of the non-Philadelphians shows a classic "Weberian" function, where the ability of the speaker to discriminate is directly related to the size of the physical difference in a linear manner. In the case of the one-step stimuli, most subjects react to this task with overt scepticism: they believe that they are being tricked by being given pairs that are in fact identical. Nevertheless, they almost all perform well above the chance level of 50%. The Philadelphians do not differ significantly from the non-Philadelphians in this one-step task. They are also not far behind the non-Philadelphians in the two-step task, wehre the difference is at the .025 level. But in the one-step task, we see a remarkable phenomenon. As the task becomes easier, the Philadelphians get worse. This is an unusual result.

As noted above, there is considerable individual variation among Philadelphians: The low value for 3-step discrimination is largely due to the performance of one quarter of the sample. We might explain the behavior of these subjects by following account. Like most other Philadelphians, they have suspended the function of semantic contrast formerly associated with /e/ and /U/ before intervocalic /r/. Two of them respond to the one-step stimuli at chance levels, but three others match the level of non-Philadelphians. Though there is much variation, they do better as a whole with the two-step task, like all other subjects. When they hear the stimuli for the three-step task, they are no doubt able to identify these with the words *ferry* and *furry*. But instead of following other subjects in doing so, their performance is inhibited by the linguistic norm that, in some deep sense, *ferry* and *furry* are 'the same'; and their performance deterioriates. The fact that two subjects show performance significantly *below* chance indicates that they have the ability to perceive the differences, but confound the labels.[13]

We can therefore conclude that the introspections of Philadelphians as a whole are not reliable indicators of their ability to distinguish the two word classes. This and many other experiments indicate that the semantic function of contrast for /e/ and /U/ is suspended in Philadelphia. The norm that *ferry* and *furry* are 'the same' intervenes between the speakers' productive system and any inquiries about it. At the same time, these norms represent a kind of linguistic reality, since /e/ and /U/ are 'the same' in this contextual position as far as their ability to distinguish words in the every-day interpretation of speech.[14]

The physical basis.

The case of /e/ and /U/ before intervocalic /r/ shares a physical characteristic with almost all other cases of near-merger. The distinction in question rests upon an F2 difference of about 200 Hz. This applies to the case of *line* and *loin* in Essex, of *sauce* and *source* in New York, of *fool* and *full* in Albuquerque, *fool* and *full* in Salt Lake City (Di Paolo 1988), of *cot* and *caught* in central Pennsylvania and elsewhere, of *too* and *two* in Norwich, and *meat* and *mate* in Belfast (Milroy and Harris 1980, Harris 1985). While 200 Hz is well above the just notice difference for F2 perception in controlled experimental situations (Flanagan 1955), linguistic stability seems to require a difference of at least 400 Hz difference. Thus the weakness of the physical basis for contrast appears to be an important conditioning factor behind the unreliability of intuition in these cases. Faber and Di Paolo (1995) show that other factors may contribute to discrimination besides F2; it is essentially the absence of a firm F1 distinction that is the major trigger for near-mergers.

The case of the Swedish vowels studied by Janson and Schulman is then even more important, since it does not depend upon an F2 difference, and indicates that social and cognitive factors can override even the robust F1 dimension.

Mismatch of intuitions and behavior in sound change: the merger of *cot* and *caughtt*

The largest phonological change taking place in the United States today is the unconditioned merger of /o/ and /oh/ in *cot* and *caught*, *Don* and *Dawn*, etc. The Telsur survey of change in progress in North America gathers information on speakers' intuitive judgments of this situation (in minimal pair tests) and their actual behavior (in minimal pair tests, elicitation, and spontaneous speech). Following Herzog's principle that mergers expand at the expense of distinctions (Herzog 1965, Labov 1964:313-4), we find that the two word classes show a strong tendency to merge in transitional areas among younger speakers. In such areas, we do not expect intuitions to match behavior. On the contrary, there is a broad and well-supported tendency for the merger to take place in perception before it takes place in production. Di Paolo's studies of the merger of *fool* and *full* in Salt Lake City (1988) and Herold's investigations of the /o/-/oh/ merger in Eastern Pennsylvania (1990) give ample documentation of this principle. Table 4 shows the relations of production and perception for the current data from the Telsur survey for the merger of /o/ and /oh/ before /t/. For all speakers, mismatches of perception and production amount to 24%. Perception leads production by better than two to one. For those under 30, the ratio is 4 to 1. This case is parallel to the near-merger of *ferry* and *furry*, in that the bias exhibited in the speakers' intuition is the result of the suspension of phonemic contrast, the first step in merger.

Table 4
Relations of perception and production
in the merger of /o/ and /oh/ before /t/

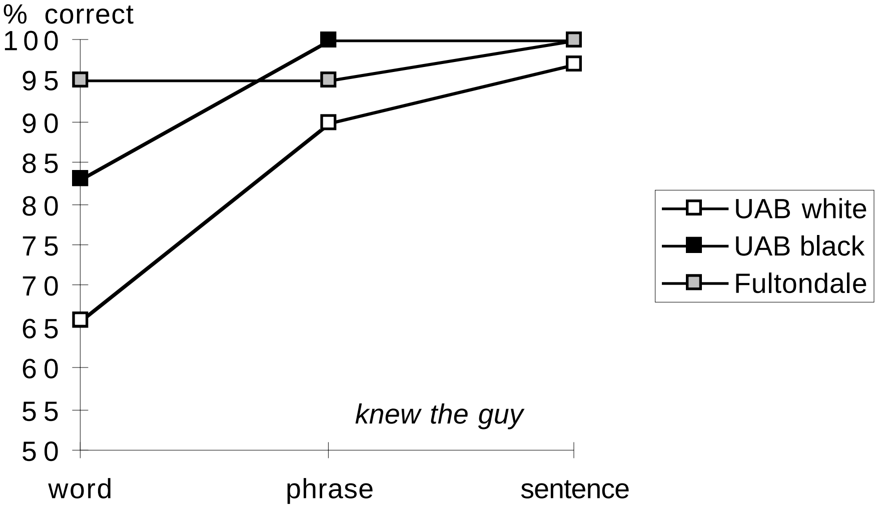
	All speakers	Under 30
Perception > production	74	20
Perception = production	330	69
Perception < production	30	5

Mismatch of intuition and production in sound change: the monophthongization of /ay/.

In the study of sound change in progress, the relations of production to perception are quite different for chain shifts and mergers.While perception is in advance of production for mergers, the situation is the reverse for ongoing chain shifts such as the Northern Cities Shift or the Southern Shift (Labov 1994: Ch.6). There is nothing resembling a minimal pair test for chain shifts, but other types of data from the studies of Cross-Dialectal Comprehension show the relationship of perception to production (Labov 1989, Labov and Ash (in press). In the Gating experiments, subjects from Birmingham, Philadelphia and Chicago listened to excerpts from spontaneous speech of speakers from their own neighborhoods as well as speakers from other cities. Subjects hear first isolated words; then phrases; then entire sentences. Local judges always showed a distinct advantage over judges from other cities, but they also had considerable difficulty in recognizing advanced phonetic forms similar to the ones they used themselves. In Chicago, the Northern Cities Shift is not recognized as such and is never a subject of overt social comment; nevertheless, only a minority of the Chicago subjects recognized isolated words using the advanced forms of the Northern Cities Shift.

In Birmingham, the Southern Shift includes some elements that are widely recognized. The monophthongization of /ay/ is a stereotype of Southern speech, and popular books on how to talk "Southern" make free reference to it. Nevertheless, Figure 10 shows only 65% success in recognizing the word [ga::] as 'guy' for 31 white college students

Figure 10. Per cent correct in identification of monophthongal /ay/ in GUY for Birmingham subjects in Cross-Dialectal Comprehension



from the University of Alabama-Birmingham. In the context of the phrase *knew the guy*,the success rate rose to almost 90%, yewt 10% of these subjects failed to recognize the clearly articulated long monophthong [ga::] as representing the word or proper name *Guy*, although this is by far the most common way of pronouncing it in Birmingham.

It is significant that two other groups of subjectsdid considerably better on this task. A small group of 6 Black students from the same college showed a higher rate of success, 83% in the word task, and 100% for the phrase task. Another group of 41 Birmingham high school students did even better in the word task, with 95%. These results lead us to the conclusion that the white college students were more open to the influence of the competing norms of
the network standard pronunciation, [ga], and itis the interference of this norm that is responsible for their inability to recognize their own speech pattern.

A general characterization of the problems.

The two problems posed at the outset were to discover the conditions under which intuitions fail, and to construct some mode of action when they do fail. So far, I have chipped away at the edges of the first problem, isolating the areas where intuition has been found to be most unreliable. We can identify five conditions that promote the failure of linguistic intuitions.

(a) *Social intervention*: when a socially superordinate norm takes precedence over the native system.

Stockholm /i~e~a/ > Lycksele /i~e~E~a/
General *ain't* as copula, pres perfect > AAVE *ain't* as preterit
General *been from (ha)s been* > AAVE Remote Pres Pf BIN
Northern/Southern *nowadays* > Midland *anymore*.
Network standard [ga] > Southern standard [ga::]

These cases reinforce the general caution sounded by Baugh (1983), that linguists who are members of a minority community should be quite wary of using introspection, since they may be biased towards or away from the literary standard in ways that can't be foreseen. On the whole, linguists are less likely to be dominated by literary norms than naive judges, but are more subject to bias from their theoretical predispositions. The difficult case here is positive *anymore*, where the social bais is not at all obvious. It is possible that any grammatical pattern that is perceived as regional may be suppressed in introspection, whether or not a social stigma can be detected.

(b) *Physical collapse*: when the physical basis for a distinction is weak or eroded
As discussed in the preceding section, most of the phonological asymmetries discussed so far depend upon the limitations of F2. It is not difficult to find grammatical correlates, as for example, the case of *could of* + past participle. Here the reduction of /hæv/ to /v/ leads to the reinterpretation of *have* as *of*. In Hawaiian Creole English, the gradual collapse of *wen* to [w] or [N] makes it difficult for even native speakers to detect (Labov 1992).

(c) *Semantic suspension*: when the semantic function of a productive distinction is suspended.

The cases of *ferry/ferry* in Philadelphia and other near-mergers
Merger in perception precedes merger in production for
/o/-/oh/ and other mergers.

The collapse of semantic distinctiveness may be a stage triggered by condition (b), but it may operates as a factor in its own right, in grammar as well as in phonology. The ongoing collapse of the perfect and preterit in German, and the accomplished collapse of the French preterit and perfect imply a stage where intuitions are compromised.

(d) *Cognitive interference*: cognitive strategies determine linguistic preferences.
NEG-Q and *NEG-V* "dialects" determined by the preference for the included rather than including interpretations.

Here we have only a single observation of what may be a major aspect of introspective confusion. All those who work with intuitions have warned against interfering cognitive factors, but little has been done to identify them and neutralize them.

(e) Pragmatic opacity: when the pragmatic function of a form is inconsistent with overt recognition by users.

G'bye > bye-bye.

Again, I have given only a single instance of what must be a widespread tendency in pragmatics. This is one of a subtype of linguistic processes that lead to instability. The strengthening of any particular linguistic signal by the multiplication of redundant signals (as in negative concord) will lead to a reduction of the semantic load carried by any one signal, and eventually to the elimination of these weakened signals. How such a process interacts with intuitions remains to be determined.

These five conditions give us a first set of guidelines that will alert us to the probability that intuitions will not correspond to behavior. Whenever one of them is recognized, it will follow that unsupported inquiries into speakers' introspections are not sufficient to decribe the state of the language. Other methods, which require more time and effort, will be required to yield an accurate view of the linguistic system. This paper has referred to a variety of modes of investigation that may be useful.

What after all can be said about the reliability of intuitions? For the great majority of sentences cited by linguists, they are reliable. In many cases, judgments that a certain linguistic form is ungrammatical may actually be motivated by the fact that it is rare, and I have no doubt that our confidence in intuitions will increase as they are coupled with the frequencies in speech production. The problem that we focus on here is to locate, specify and explain the minority of judgments that are dramtically different from the data on frequency of speech production. Until we can do so, any given linguistic theory may be incoporate unknowingly data that appear to support but actually controvert the principle being asserted.

Whenever there is a physical basis for the unreliability of judgments, we are on firm ground. Thus we can predict with certainty that any phonological distinctions that are supported by mean differences of 100 to 200 Hz F2 in production will not be accompanied by reliable judgments of 'same' or 'different'. So far, there are no counterexamples to this principle.

In the case of the northern Swedish dialects, there is no such physical basis, but it seems clear that the standard norm of Stockholm replaces the local Lycksele norm when categorization judgments are made. But this is a post-hoc observation. There are other northern Swedish dialects that maintain four vowels, where judgments are not subverted by the Stockholm norm. Until an explanation of this difference is brought forward, we can not predict when the standard norm will have this effect. Another strategy would be to call into question judgments on any local or regional pattern. Though this would prevent us from being deceived by a Lycksele-type response, it would slow down our work enormously with the great majority of local and regional dialects that are not so influenced.

For phonological studies, it is relatively easy to couple intuitive judgments with observations of behavior; it is not so simple with the less frequent syntactic and morphological variables. We have yet to understand why judgments on positive *anymore* are so skewd from reality, while judgments of *needs washed,* etc., seem to be much more robust.

If the social pressures on a minority or local community are well recognized, we can employ considerable caution in interpreting intuitive judgments. But this does not explain why judgments about *ain't* in the preterit *ain't* are so highly skewed, while judgments about *bin* are less so, and the recognition of invariant *be* seems to be quite straightforward. While frequency of forms plays some role, the in the negative preterit shows a very high frequency indeed.

Until such problems are resolved, all linguists should be cautioned that the use of intuitive judgments of acceptability as the sole basis for linguistic generalization may incorporate serious errors into the argument. Attention to the patterns of spontaneous speech, on the most informal basis, may help to indicate where these errors lie.

Notes

[1] See for example, Grinder and Postal 1971, Lakoff 1973.

[2] The one subject who did not later showed a NEG-V reaction to the syllogistic question: "All men do not have three arms; John is a man; therefore John _____."

[3] Difference between object and subject interpretation: $c^2 = 21.03$, $p<.001$ (given with the Yates correction).

[4] In the field of phonology, there have been a number of recent efforts to modify the discrete character of optimality theory to deal with gradient phenomena (Nagy and Reynolds in press, Zubritskaya in press).

[1] Observation made by Teresa Labov

[5] Observation made by Barbara Freed, who conducted this interview.

[6] In the Eastern part of the Midland area, positive *anymore* associated strongly with a negative social evaluation, that is, a complaint. When Guy Lowman first began noting positive *anymore* sentences in his notebooks in the course of his Linguistic Atlas interviews in Pennsylvania and West Virginia, he began with a neutral sentence, and got variable responses. When he switched to *Farmers are pretty scarce around here anymore* he got far more consistent responses.

[7] In some cases, typically a game, the time may be only a matter of seconds, but it is asserted to be long from a competitive point of view.

[8] T students carrying out the survey were not familiar with AAVE BIN. They were drilled in the pronunciation of BIN with a low, stressed tone, but it cannot be shown how successful this training was. But any effect of incorrect pronunciation by the student administrators of the test would not explain the high percentage of "yes" responses by whites with strong black contacts.

[9] In more recent studies in Philadephia, the use of *ain't* rises to 70%.

[10] These are only the most common of the items studied. The full data set shows 23 different surface forms, including a number of phonetic variants of *bye bye* and *goodbye*.

[11] Categorization of the barks was considerably enhanced by the shortness of the stimuli, and it is thetefore difficult to compare them to the other tasks; they will not be considered further in this discussion.

[12] In a preliminary investigation of the ability of speakers from Toronto to hear the *cot/caught* distinction, Herold found a minority in this merged dialect who showed similar behavior: they could label the Mid-Atlantic distinction consistently, but reversed the labels.

[131] See the Coach Test in Labov 1994: Chapter 14: 404 ff.

References

Bard, Ellen, Dan Robertson and Antonella Sorace 1996. Magnitude estimation of linguistic acceptability. Language 72.1-31.

Baugh, John. 1983. Black Street Speech: its history, structure and survival. Austin: University of Texas Press.

Carden, Guy. 1970. A note on conflicting idiolects. Linguistic Inquiry 1:281-290.

Chomsky, Noam. 1981. Lectures on Government and Binding. Dordrecht, Netherlands: Foris Publications.

Dayton, Elizabeth. 1996. Tense and aspect in the Philadelphia black community. University of Pennsylvania dissertation.

Di Paolo, Marianna 1988. Pronunciation and categorization in sound change. Linguistic Change and Contact: NWAV XVI, ed. by K. Ferrara et al., 84-92. Austin, TE: Dept of Linguistics, U. of Texas.

Di Paolo, Marianna and Alice Faber. 1990. Phonation Differences and the phonetic content of the tense-lax contrast in Utah English. Language Variation and Change 2:155-204..

Faber, Alice and Marianna Di Paolo. 1995. The discriminability of nearly merged sounds. Languge Variation and Change 7:35-78.

Flanagan, J.. 1955. A difference limen for vowel formant frequency. JASA 27:613-617.

Grinder, John and Paul Postal 1971. Missing antecedents. Linguistic Inquiry 2:209-312.

Harris, John. 1985. Phonological variation and change: studies in Hiberno-Irish. Cambridge, MA: U. of Cambridge Press.

Heringer, James T.. 1970. Research on quantifier-negative idiolects. CLS 6:287-96.

Herold, Ruth. 1990. Mechanisms of merger: The implementation and distribution of the low back merger in Eastern Pennsylvania. U. of Pennsylvania dissertation.

Herzog, Marvin I.. 1965. The Yiddish Language in Northern Poland. Bloomington & The Hague. (= IJAL 31.2, Part 2).

Hindle, Donald 1974. Synteactic variation in Philadelphia: Positive anymore. Pennsylvania Working Papers on LInguitic Change and Variation, No. 5. Philadelphia: Linguistics Laboratory University of Pennsylvania,

Janson, Tore and Richard Schulman. 1983. Non-distinctive features and their use. Journal of Linguistics 19:321-336.

Karins, Krisjanis and Naomi Nagy. To appear. Testing the perception of a "categorical" rule: Wanna experiment in syntax? Linguistic Variation and Change.

Labov, William 1994. Principles of Linguistic Change. Volume 1: Internal factors. Oxford: Blackwell Publishers.

Labov, William in press. Co-existing systems in Afrian American Vernacular English. *The Structure of African-American English*, ed. byS. Mufwene, J. Rickford, J. Baugh and G. Bailey.

Labov, William and Sharon Ash. To appear. Understanding Birmingham. In C. Bernstein, R. Sabino and T. Nunnally (eds.), Proceedings of LAVIS II.

Labov, William, Malcah Yaeger & Richard Steiner. 1972. A Quantitative Study of Sound Change in Progress. Philadelphia: U. S. Regional Survey.

Labov, William, Mark Karan and Corey Miller. 1991. Near-mergers and the suspension of phonemic contrast. LVC 3:33-74.

Labov, William 1966. The Social Stratification of English in New York City. Washington D.C.: Center for Applied Linguistics.

Labov, William. 1972. Where do grammars stop? Georgetown Monograph on Languages and Ling 25, ed. by R. Shuy, 43-88.

Labov, William. 1975. What is a linguistic fact? Lisse: Peter de Ridder Press. Lisse: Peter de Ridder Press. Also as Empirical foundations of linguistic theory. The Scope of American Linguistics, ed. by R. Austerlitz, 77-113. Lisse: The Peter de Ridder Press.

Labov, William. 1989a. The exact description of the speech community: short a in Philadelphia. Language Change and Variation, ed. by R. Fasold and D. Schiffrin, 1-57.. Washington, Georgetown U.Press

Labov, William. 1989b. The limitations of context. CLS 25, Part 2, 171-200.

Labov, William. 1992. Onthe adequacy of natural languages I: the development of tense. Pidgin and Creole Tense-Mood-Aspect Systems, ed. by J. Singler, 1-58. Amsterdam/Philadelphia: John Benjamins.

Lakoff, George 1973. Fuzzy terminology and the performance/competence game. CLS 1973:271-91.

Milroy, James and John Harris. 1980. When is a merger not a merger? the MEAT/MATE problem in a present-day English vernacular. English World-Wide 1:199-210.

Nagy, Naomi and Bill Reynolds in press. Optimality theory and variable word-final deletion in Faetar. To appear in Language Variation and Change.

Newmeyer, Frederick 1983. Grammatical theory: its limits and its possibilities Chicago : University of Chicago Press, 1983

Nunberg, Geoffrey. 1980. A falsely reported merger in eighteenth century English: a study in diachronic variation. Locating Language in Time and Space, ed. by W. Labov, 221-250. New York: Academic Press.

Schütze, Carson T. 1996. The Empirical Base of Linguistics: Grammatical Judfgments and Linguistic Methodology. Chicago: U. of Chicago Press.

Voegelin, C. F. and Harris, Zellig S.. 1951. Methods for determining intelligibility among dialects of natural languages. Proceedings of the American Philosophical Society 95:322-329.

Zubritskaya, Katya in press. Sound change in OT. To appear in Language Variation and Change.

[1] See for example, Grinder and Postal 1971, Lakoff 1973.

[2] The one subject who did not later showed a NEG-V reaction to the syllogistic question: "All men do not have three arms; John is a man; therefore John _____."

[3] Difference between object and subject interpretation: $c^2 = 21.03$, $p<.001$ (given with the Yates correction).

[4] In the field of phonology, there have been a number of recent efforts to modify the discrete character of optimality theory to deal with gradient phenomena (Nagy and Reynolds in press, Zubritskaya in press).

[5] Observation made by Teresa Labov

[6] Observation made by Barbara Freed, who conducted this interview.

[7] In the Eastern part of the Midland area, positive *anymore* is associated strongly with a negative social evaluation, that is, a complaint. When Guy Lowman first began noting positive *anymore* sentences in his notebooks in the course of his Linguistic Atlas interviews in Pennsylvania and West Virginia, he began with a neutral sentence, and got variable responses. When he switched to *Farmers are pretty scarce around here anymore* he got far more consistent responses.

[8] In some cases, typically a game, the time may be only a matter of seconds, but it is asserted to be long from a competitive point of view.

[9] T students carrying out the survey were not familiar with AAVE BIN. They were drilled in the pronunciation of BIN with a low, stressed tone, but it cannot be shown how successful this training was. But any effect of incorrect pronunciation by the student administrators of the test would not explain the high percentage of "yes" responses by whites with strong black contacts.

[10] In more recent studies in Philadephia, the use of *ain't* rises to 70%.

[11] These are only the most common of the items studied. The full data set shows 23 different surface forms, including a number of phonetic variants of *bye bye* and *goodbye*.

[12] Categorization of the barks was considerably enhanced by the shortness of the stimuli, and it is thetefore difficult to compare them to the other tasks; they will not be considered further in this discussion.

[13] In a preliminary investigation of the ability of speakers from Toronto to hear the *cot/caught* distinction, Herold found a minority in this merged dialect who showed similar behavior: they could label the Mid-Atlantic distinction consistently, but reversed the labels.

[141414] See the Coach Test in Labov 1994: Chapter 14: 404 ff.