The Reliability of Linguistic Testimony about "Staged" vs. Natural Conversation: Perspectives of Linguistics and Law

George Castelle, J.D.,
Chief Public Defender,
Charleston, West Virginia,
and
Roger W. Shuy, Ph.D.,
Georgetown University
Washington, DC

Abstract

In the past decade, U.S. courts have struggled with the admissibility of expert testimony in instances where established forensic science techniques have been applied to new and unfamiliar legal disputes. This controversy is particularly pronounced in fields such as forensic linguistics. The authors examine the issue of whether recorded conversations by criminal suspects are natural or staged, and whether a forensic linguist's analysis of these conversations should be admissible in U.S. courts. The linguistic analysis involves a comparison of natural versus staged speech, including such distinguishing factors as the frequency of overlap, interruptions, topic recycling, contractions, pause fillers, and intonation patterns. The legal analysis identifies misapplications of the court's reliability assessment, as required by the United States Supreme Court in Daubert v. Merrell Dow Pharmaceuticals, particularly in instances where a narrow and inflexible application of Daubert criteria result in the inadmissibility of sound and valuable expert testimony.

Key Words

Forensic science, forensic linguistics, staged conversation, expert witnesses, Daubert hearings, reliability assessment

In recent years lawyers have called on linguists more and more to assist them with their cases. This makes good sense, since legal issues are often about language. Although lawyers have great expertise in using language, they often

are not experts in how language actually works. During the same period, the field of law has been reassessing how to deal with experts in general and the knowledge they can bring to their cases. This paper is the combined effort of a linguist and a lawyer to focus on one of the problems that this new legal thinking has produced. This problem associates with how to evaluate the recently required "reliability" of expert witness testimony. The paper is in two parts. The first part is written by a linguist who has served as an expert witness in some fifty cases. The second part is by a lawyer who has represented clients in hundreds of cases.

THE LINGUISTIC PERSPECTIVE

One of the most interesting aspects of working with language problems in the world of law is that law provides questions that linguists never think to ask. For example, attorneys in a civil case once asked a linguist to try to determine whether a client's notes were made during a conversation, after it, or from a secret tape of that conversation (1). More recently linguists have been asked another question that, as far as we know, has not been well researched: "Were certain tapes that were in evidence in a criminal case naturally occurring conversation or were they staged by one or more participants who conspired to make the tapes, like actors, for later listeners to hear?" In other words, were these taped conversations fake or real? Were they performed from prearranged scripts rather than occurring spontaneously?

Two recent criminal cases illustrate this issue: United States v. Shalom Weiss (bank fraud) (2) and Tennessee v. Byron Looper (murder) (3). In both cases, tape recordings of conversations were presented to the Court. In one case, the defense claimed the tapes were staged. In the other case the prosecution made the same accusation.

Background of United States v. Weiss

The tapes in evidence in United States v. Weiss (2) were a small part of a very large case involving bank fraud, with multiple defendants and complex financial and language components. One defendant, somewhat marginal to the case as a whole, was charged with a small part of the huge overall fraud. The major defendants had begun to plead guilty and then proceeded to surreptitiously tape other parties as part of their plea agreement. As is common in such situations, it behooves the cooperating witnesses to capture as many others as possible, for such action helps reduce their own sentences. This activity sometimes works but sometimes leads law enforcement down unproductive paths. Occasionally it also nets people who, though not guilty of the charges, are made to appear guilty by the way they are taped, thus producing an opportunity for linguists to try to untangle the mess created, among other things, by the ambiguity and vagueness of the conversations.

In this case the cooperating witnesses taped this defendant several times, with results that were, at best ambiguous. After a while, the defendant became suspicious about the way these conversations were going, and decided to himself tape record the cooperating witnesses whenever they came by. A central tape in this case, however, was produced by the cooperating witness who claimed that its exculpatory nature was contrived, a result of both speakers staging a tape which, if they were caught, would work toward establishing their innocence.

Naturally, the defense chose to play this tape to counter the cooperating witnesses allegations that defendant Weiss was guilty. Therefore, when the cooperating witnesses took the stand as a prosecution witness against Weiss, the

defense confronted him on cross examination with these two tapes. He at first shrugged them off as compilations of fragmented pieces of conversation spliced together for appearances sake. To rebut this claim, an acoustic engineer, brought in by the defense, testified that the tapes were not spliced and not altered in any way. Hearing this, the cooperating witness then changed his story and testified that the tapes were staged by himself and the defendant, and the two of them made these tapes to create the appearance of innocence. His testimony was that our defendant was, indeed, as guilty of the scheme to defraud as he himself was.

The defense attorney was then placed in the position of needing to counter the cooperating witness's claim that the tapes were staged. At this point, the linguist co-author here was called upon to try to determine whether these tapes were staged or, as the defense claimed, naturally occurring conversations.

Definitions of staged and natural

It was first necessary to define what "staged" might mean in this context. It is clear that it must refer to conversations in which both parties were aware that a tape is being made and who used this awareness to produce, together and in consort, a conversation that had as its goal the presentation of information favorable to a certain position and unfavorable to other interpretations. The resulting language of a staged tape differs from that found in conversations in which neither of the parties knows that a tape is being made, in which case the resulting language is natural and not self-consciously produced. In contrast, the participants in staged conversations, who are mutually constructing the event, are very conscious of their planned roles and are careful not to step on each other's lines, not to interrupt each other, and not to disagree with each other, since these things could cause later listeners to be confused or unclear about what they are saying. Staged tape participants are also more careful to complete their sentences and topics, and to articulate loudly and carefully so that the later listeners for whom the tape is being made can understand easily and fully. A tape recorded product of staged conversation might well sound more like a stage play than a natural, everyday conversation.

There were many, many tapes in this case. Some were wiretaps, where neither party was aware that a tape was being made. Others were made by one of the parties while the other party did not know he was being recorded. In such cases, the above definition of "staged" does not apply because this requires the cooperation of the two parties to work together, in consort, to produce the results. In tape recorded undercover operations, when only one person knows he is being taped, there is usually no mutual construction of the conversation and no conscious awareness by one of the parties about the goals of the other. This often contributes to conversational spontaneity and unpredictability, which in turn encourages natural language use.

In undercover taped conversations, when only one of the participants is aware of the taping, even the person doing the taping sometimes tends to forget that he is making a tape and often reacts with natural rather than planned language use. On such tapes, the taper may preserve his conversational goals, similar to that of an interviewer's preserving his planned questions, but the resulting language is usually natural rather than consciously staged. Symmetry of naturalness in conversation is an important aspect of being cooperative. If one person uses consciously careful and overly articulate speech while the other does not, a conversational awkwardness can develop, leading to distrust on the part of the unknowing participant. In fact, it is not uncommon, when such

awkwardness occurs in surreptitious tapes, for the alert but unknowing party to then ask the taping party whether or not he is tape recording that conversation.

Factors that accompany natural conversation

With these definitions, the next step was to make a list of the characteristics of natural conversation and to examine the two tapes for evidence of whether or not these factors were present. The list included the following factors which were selected because they provide basic language differences between, on the one hand, the carefully, consciously, and cooperatively constructed dialogue of two people who know they are producing a tape for later listeners to hear, and the less carefully and consciously produced natural conversation of two participants who are not working together to produce a tape for the benefit of later listeners.

1. discourse factors

Overlapping of turns works against a mutually constructed purpose of being clear and understandable. One can distinguish overlapping from interruption in that in the latter, while both speakers talk at the same time, one of them wins out and breaks the continuity of the other's speech. This is often also accompanied by repetition of some or all of the words uttered during the overlap. Overlapping, as defined here, differs from interruption in that, although two people briefly talk at the same time, the nature of the speech is such that the speaker holding the floor is not stopped from the continuous turn.

Topic recycling can be expected to be less necessary in staged speech than in natural conversation. When the conversation is mutually and consciously constructed, the focus is on completing and/or resolving the topics so that they will be clear and understandable to later listeners. Therefore, there is little or no need to recycle topics in conversations that are staged. Speakers complete the topic, then go on to the next planned topic.

Staged conversations will obviously attempt to avoid vagueness and ambiguity. By contrast, in natural conversations, vagueness is common, primarily because the participants mutually understand what they are talking about and make references that outsiders to the understanding might not comprehend.

In natural conversation, speakers sometimes do not finish their sentences or complete their ideas and topics. Again, this is because they know (and/or see) each other well enough to assume that the other person gets their point anyway. This, of course, is not true when the conversation is intended and staged primarily for the benefit of outsiders, for whom sentences and topics tend to be completed.

Since the intent of a mutually staged conversation is to present a unified front for the benefit of later listeners, one can expect those who stage it to have few if any disagreements, challenges, or accusations between them. Such would work against the purpose of presenting a mutual representation of information favorable to the position they espouse and want later listeners to understand.

2. articulation factors

In mutually staged conversations, one can expect carefully constructed speech that strives to present a position clearly and understandably so that

even a later audience will understand it. It therefore will contain clear and audible articulation. Although the register may remain close to casual, it often moves toward formal talk. Natural conversation more commonly adheres to the casual register, with many contracted verb forms, clichÈs, obscenities, sentence fillers, feedback markers, and a wider variation in intonation that is common to everyday talk. Staged conversation can be expected to have fewer of these characteristic casual articulations.

summary of comparison factors

Discourse Structure

Natural conversation	Staged conversation
sometimes inappropriate turn taking with overlapped speech	precise turn taking and lack of overlapped speech
interruption	no interruption
topic recycling	little or no topic recycling
vagueness or imprecision	absence of vagueness and imprecision
fragmented ideas and topics	unified ideas and topics
disagreements with each other	no disagreements
accusations of each other	no accusations
sarcasm to each other	no sarcasm

Articulation Structure

Natural conversation	Staged conversation
many contractions	fewer contractions
clichÈs and obscenities	fewer clichÈs and obscenities
simple syntax	simple plus complex syntax
many sentence fillers	fewer sentence fillers
many feedback markers	fewer feedback markers
varied intonation patterns	flattened intonation or even inappropriate intonation

In should be pointed out that although these factors commonly distinguish natural from staged conversation, they cannot, at this point, be said to reach the level of being called completely distinctive features, since the defining

research necessary to distinguish staged from natural conversation has not yet been done. This will be seen as a crucial aspect in the Weiss case.

The Proposed Testimony: Comparing the Unstaged with the Allegedly Staged Tapes in the Weiss case

Since there is an absence of research on how people stage conversations, the logical procedure to follow in this case was to compare the language of the tapes alleged to be staged with other tapes of the same speakers but for which there is agreement that they were natural and unstaged. It would have been optimal for such a comparison if the tapes in evidence contained a wiretap conversation between the same participants. Unfortunately, this was not the case. The wiretap tapes in this case were between different defendants. The only possible comparison tapes were those in which one of the parties knowingly taped the other, unknowing party. There was only one such tape that contained the defendant and one of the cooperating witnesses.

In short, two tapes were alleged to have been staged between the defendant (D) and one of the cooperating witnesses (CW-1). As examples of tapes for which there was agreement by both the prosecution and the defense that there was no staging involved, two comparison tapes were offered, one between the defendant (D) and that same cooperating witness (CW-1) and the other between the defendant (D) and a different cooperating witness (CW-2), who did not appear on the two allegedly staged tapes. The latter tape (between D and CW-2) was not the perfect match for comparing the allegedly staged interaction, but it did serve to show how D spoke under natural conditions. Experimental research might be able to come up with better comparison tapes, but such conditions do not often exist in the real world (of which more will be said later).

The tapes were not equal in length, but similar enough for comparative purposes. The lengths, excluding telephone interruptions that occurred during the admittedly unstaged meetings, were as follows:

Unstaged tape-1	19:08	
Unstaged tape-2	30:13	
Alleged staged tape-1		22:22
Alleged staged tape-2		7:51

Although the total length of the two known unstaged tapes is a third longer than the length of the two alleged staged tapes, the number of turn exchanges in both the unstaged and allegedly staged tapes is almost identical, 145 to 144. This difference can be accounted for by dissimilar tasks and topics. In the known unstaged tapes, the participants are going through various documents together and have longer pauses between their turns of talk.

The units of analysis for comparing the two sets of taped conversations are topic and turn exchange, as the following will demonstrate.

The language of the known unstaged tapes

The first task was to determine the extent to which the language features of natural conversation, listed above, were present in the known unstaged tapes, UST-1 and UST-2, as follows:

Feature	UST-1	UST-2
overlapping	Of the 45 turns of talk, 20 contained overlaps, a rate of 44%	Of the 100 turns of talk, 33 contained overlaps, a rate of 33%
interruption	4 interruptions, a rate of 10%	10 interruptions, a rate of 10%
topic recycling	6 of 13 topics recycled, a rate of 41%	15 of 28 topics re-cycled, a rate of 53%
vagueness	7 vague references	7 vague references
fragmentation	4 fragment utterances	4 fragment utterances
disagreements	3 disagreements	4 disagreements
challenges	none	none
accusations	none	none
sarcasm	none	none
contractions	16 forms, multiply used	16 forms, multiply used
clichÈs	2 cliches	1 cliche
obscenities	2 obscenities	3 obscenities
syntax	mostly simple sentences	mostly simple sentences
sent. fillers	"you know," "I mean"	"okay," "alright"
intonation	wide variation	wide variation

It should be noted that in the first of the two known unstaged conversations (UST-1), the defendant and the cooperating witness (CW-1) discussed financial resources and how to pay bills. In the second known unstaged conversation, the defendant and the cooperating witness (CW-2) talked while they went through various papers trying to discover the identity of various business entities. In both cases, the pace of conversation was rather slow and the emotional tone rather calm. The participants were working in consort and appear to be agreeable to each other, as least as far as this task was concerned.

The language of the alleged staged tapes

Next the alleged staged tapes (AST-1 and AST-2) were analyzed to determine the presence or absence of the above language features:

Feature	AST-1	AST-2
overlapping	Of the 107 turns of talk, 83 contained overlaps, a rate of 77%	Of the 37 turns, 29 were overlapped, a rate of 78%
interruption	26 interruptions, a rate of 24% per turn	10 interruptions, a rate of 27% per turn
topic recycling	22 of 45 topics recycled, a rate of 48%	8 of 19 topics recycled, a rate of 42%
vagueness	14 vague reference	4 vague references
fragmentation	5 fragment utterances	2 fragment utterances
disagreements	7 disagreements	4 disagreements
challenges	6 challenges	5 challenges
accusations	1 accusation	6 accusations
sarcasm	1 sarcastic remark	2 sarcastic remarks
contractions	20 forms, multiply used	10 forms, multiply used
clichÈs	5 cliches	1 cliche
obscenities	7 multiple obscenities	2 obscenities
syntax	mostly simple	mostly simple
sent. fillers	"you know," "I mean"	"you know," "I mean"
intonation	wide variation	wide variation

It should be noted that in both of these allegedly staged conversations, the defendant was trying to find out what the cooperating witness (CW-1) knew about the financial difficulty that he faced and even any possible legal difficulty that he might encounter. The defendant clearly suspected that CW-1 knew considerably more than he was telling. He challenged, accused, and used sarcasm. Both parties disagreed with each other several times. This conversation was considerably more animated and fast paced than it was in the two known unstaged tapes. Other than this, however, the two alleged staged tapes were consistently similar to the known unstaged tapes in their uses of these language factors.

Overlapping and interruption actually occurred at higher rates in the alleged staged tapes than in the known staged one. Lower rates of overlap and interruption might have been expected in the allegedly staged conversations. It is difficult to imagine how a staged conversation might cause high rates of overlap and interruption since these factors create less audible and clear language and are counterproductive to the creation of a comprehensible tape for the benefit of later listeners.

What we find here is the exact reverse of the degree of overlapping and interruption that one might expect: high rates in the allegedly staged conversations and lower rates in the known unstaged ones. The lower rates of overlap and interruption in the known unstaged tapes are no doubt a result of the topics being discussed. In the known unstaged tapes, the participants were doing exactly what a staged tape might produce--a mutually cooperative effort, in this case, trying to find answers to their questions about financing and the identity of various financial entities. In contrast, in the alleged staged tapes, the participants were not working in consort, as staging would predict, but were at severe odds with each other, hardly conducive to any interpretation that they were staging the events.

That topic recycling will occur more commonly in natural conversation than in consciously produced or staged conversations is common sense (But once again, we must note that there is no available empirical evidence to support this common sense view).

Although there is no benefit from recycling in a staged conversation since the purpose is to make points in a clear and unified manner, the known unstaged tapes in this case recycle topics at a rate quite similar to that of the alleged staged tapes.

Although common sense tells us that it will take longer for speakers in a staged tape than an unstaged one to introduce topics, make them clear and explicit, and conclude them, once again we have no empirical, published evidence to support this. In this case, the average elapsed time for each topic in the alleged staged tapes was one topic every 47 seconds. The average elapsed time for each topic in the known unstaged tapes was 107 seconds. This difference in elapsed time per topic is noteworthy, since a staged tape could be expected to produce more time per topic as the participants make clear and unambiguous the points they want later listeners to understand. As it turns out here, the alleged staged tapes are actually a minute shorter per topic than the known unstaged ones, a point which runs counter to what can be expected in a staged conversation.

The 12% frequency of vague references per turn exchange in the alleged staged conversation (18 vagueness in 144 turns) is very close to the 9% frequency in the known unstaged conversations (14 vaguenesses in 144 turn exchanges). This 3% difference is significant in that one would expect a lower rate of vagueness in staged conversations than in unstaged one. That the rate of vagueness is higher in the alleged staged conversations than it is in the known unstaged ones suggests strongly that these tapes are not staged.

The rate of occurrence of fragmented sentences is virtually identical in the known unstaged and allegedly staged conversations. There are 8 fragments in 145 turn exchanges in the unstaged tapes (5%) and 7 fragments in 144 turn exchanges in the alleged staged conversations (5%). Again, common sense tells us that unstaged conversation will have a higher frequency of fragments than a planned, staged conversation would produce.

Frequency of the negative speech acts of disagreement, challenge, and accusation differ greatly between the known staged and alleged unstaged conversations. In the alleged staged tapes, 29 of the 144 turn exchanges (20%)

contain these features. In the known unstaged tapes, they occur in only 7 of the 145 turn exchanges (5%). Once again, however, the expectation of an alleged staging would predict just the reverse. Such negative speech acts do not promote the construction of a unified position common to staging. The high rate of occurrence of these features in the alleged staged tapes suggests strongly that these tapes are not staged, as does the use of sarcastic remarks to each other, as we find in the alleged staged tapes but not in the known unstaged ones.

The articulation features of the known unstaged and alleged staged tapes are substantially similar, once again counter to common sense expectation. Both sets of conversations are in the casual register, never veering toward the more formal articulations expected for a staged tape. Both sets use contracted verbs most of the time. Both are liberally sprinkled with clichÈs and obscenities. Both lean on simple sentence structure with occasional compounding but rarely contain any subordination. Both use many sentence fillers such as "I know," "okay," and "I mean." Finally, the intonation of both sets of tapes is consistently natural, with rising and falling intonation in appropriate places. In a staged conversation, one expects a veering off into stage-like intonation, uncontracted verbs, more complex syntax reflecting greater thoughtfulness, and fewer sentence fillers, all of which suggest strongly that the alleged staged tapes are not staged after all.

Objections to the Proposed Linguistic Testimony

It was noted earlier that the real world conditions producing the context in which this comparison was made are not the perhaps ideal experimental ones. When the above analysis was presented during voir dire, the prosecutor objected strongly, observing that the proposed linguistic testimony was not "scientific knowledge." She quoted from the United States Supreme Court case of Daubert v. Merrell Dow Pharmaceuticals (4) as follows:

> ...in order to qualify as 'scientific knowledge, an inference or an assertion must be derived by the scientific method. Proposed testimony must be supported by appropriate validation--i.e., 'good grounds,' based on what is known.

The prosecutor then went on to question whether or not the conclusions reached were based on experimental tests that validate these findings. There was no objection about qualification as an expert or about linguistics being anything but a legitimate science. Her only point was that the proposed testimony, using linguistic features to determine whether or not a tape was staged or natural, was not based on a preceding laboratory comparison. The notion of scientific validation, therefore, was narrowly defined as experimentally based.

Of course, no evidence could be offered that the proposed comparison was preceded by or based on previous laboratory experiments. It was considered of no consequence that naturally occurring language would no longer be naturally occurring once it was placed in the experimental laboratory and that any research of the type she suggested (she admitted to being a psychology major in college) would be impossible for this type of science. Even if laboratory subjects were given samples of both staged and natural conversation to judge, the results, whatever they might be, would result only in subjective reactions to language and would ignore the actual structure of the language evidence used as evidence. It was further suggested that there are many types of science

besides experimental, as is found in the practice of much of anthropology, linguistics, and even biology. Whenever the human factor is studied, the experimental, laboratory approach can actually distort the findings rather than aid them.

This argument did not prevail, and the judge ruled against permitting the proposed testimony before the jury. This is a situation that linguists who prepare to testify at trial must consider and address. In the US. at least, the new law on the admissibility of expert witnessing by linguists , based on "reliability," is not kind to the type of science that many experts, including linguists, commonly practice. This is especially problematic for linguistic questions such as whether or not a conversation was staged or natural, which have not been previously considered or addressed by linguists. There has to be a first analysis when none exist previously. Does this doom science to only a reliance on past research? Does it exclude all science but that which is experimentally based?

It has been suggested that in cases such as this, linguists should be clear that their task is not to state a conclusion about whether or not the tapes were staged but, rather, to educate the jury about the differences between staged and unstaged conversations (5). Solan (6) also makes this very point, referring to one important role of the linguist as a "tour guide" for the jury, helping them wrestle with issues that they had hitherto not considered. However much there is to be said for such a position, the above effort to follow this advice was never given an opportunity for testing.

The stark facts of Daubert remain an obstacle for linguistic expert witness testimony. The Daubert standard was supposed to loosen requirements for expert testimony, but that is not how it worked out in this case. The listing of specific factors that a judge can take into account may well make admission less likely in the future. It is hoped that this analysis and experience will be instructive to forensic linguists in approaching similar cases.

Background of Tennessee v. Looper

The question of whether or not a tape was staged was asked again, this time not by the attorneys involved, but by producers of the television series, American Justice. In the criminal case of Tennessee v. Byron Looper (3), the defendant was accused of murdering his political opponent for a seat in the Tennessee legislature, incumbent Tommy Burks. Looper, who had been the Property Assessor for Putnam County, was indicted on fourteen counts of theft and misuse of office some seven months before the election. Looper had felt animosity between himself and Attorney General Bill Gibson for some time and he stated that his indictment was just one more bit of evidence that Gibson was out to get him.

After Senator Burks was killed, Looper became the major suspect. He disappeared immediately, fanning suspicions created by various witnesses who associated him with the murder. His trial began in August, 2000. The defense tried to introduce into evidence a tape recording that Looper claimed was between Gibson and an informant who Gibson allegedly used in some cases. In that tape, the voice that was claimed to be Gibson's is heard to say: "I want you to set up Byron Looper." The alleged informant (the other speaker) on the tape refused to verify the tape's authenticity, so the judge excluded it as evidence in the trial.

Linguistic Analysis of the tape

When American Justice decided to use the Looper case for one its weekly programs, the producers decided to investigate the tape further. They wanted a linguistic analysis of it, primarily to determine whether or not the voice was actually that of Attorney General Bill Gibson. They provided a known sample of Gibson's speech for comparison purposes. Three types of linguistic comparisons were made: phonetic, lexical and discourse features. The known sample of Gibson's voice differed from that of the voice on the alleged conspiracy tape in pronunciations of the vowels of "bite," "mean," "set," "just," "Byron," the number of syllables in "Byron," voice quality (Gibson has what is called a creaky voice), and tempo. The major lexical marker was their contrastive use of the variables, "yeah," "uh-huh," and "yes." In terms of discourse, half of Gibson's sentences begin with the discourse marker, "Well," while the conspiracy- tape speaker used none. Gibson repeats words 92 times, especially the words, "and," "to," and "the"; the conspiracy tape speaker repeated no words at all.

From this analysis it was very clear that Bill Gibson's voice was not that of the speaker on the alleged conspiracy tape. More than this, however, the linguistic evidence supported the probability that the entire tape was staged. The carefully articulated, non overlapped turns of talk, the abnormal intonation, and the textual illogic supported this hypothesis. The following passage is illustrative:

Speaker 1: John, I got a problem I need you to take care of.

(3 second pause)

Speaker 2: What's that?

(1 second pause)

Speaker 1: I want you to set up Byron Looper.

(1 second pause)

Speaker 2: What do you mean, like what I did in Crossville?

(2 second pause)

Speaker 1: Yes, you know what I mean. Whatever you got to do, just get it done. Whatever it takes.

In unstaged conversation about a tense topic such as this, there is little likelihood of such pause lengths between turns of talk. Even one second pauses are long in most conversation. There is also some likelihood that there would be some overlap of speech. This result sounds more like speaking from a prepared script, where the speakers are being careful not to step on each other's lines.

The intonation of the exchanges is also odd. Speaker #2's "What's that?" is spoken with flat intonation, very unlike what one might expect. Speaker #1's following statement, "I want you to set up Byron Looper," puts primary stress on "I" and "Looper" where one might expect it to be on "you," "set up," and on both the first and last names of "Byron Looper" rather than on "Byron" only. Speaker #2 puts no stress on "mean" in his next turn of talk, emphasizing only

"Crossville" in that otherwise flat intonation sentence. Then Speaker #1 emphasizes "mean" rather than "know" in the following sentence, nor does he emphasize "takes." On the whole, the intonation sounded as though it was badly read from a prepared script.

The exchange about what Speaker #1 "means" is also odd.Speaker #2 asks, "What do you mean?" He then adds that it might be like what he did in Crossville. To this, the alleged Gibson says, "Yes, you know what I mean." One would expect him to have agreed by saying something like, "That's exactly right" or "Right." Speaker #2 gives evidence that he understands what the alleged Gibson means, so there is little reason for a reply. If Speaker #2's guess about it being like Crossville was wrong, the alleged Gibson's response might have more credence, since "You know what I mean" is said when the other party has not given evidence to that point that he has understood correctly.

Thus it became clear that not only was the tape not of Attorney General Bill Gibson's voice, but was also a staged performance by two different, unknown people. As noted above, the tape was not admitted into evidence and, unlike the Weiss case, the linguistic analysis did not reach the point of being accepted or rejected as testimony.

Thus a question that linguists hitherto had not viewed as important or interesting enough to merit independent research suddenly became important in two recent criminal trials. Possible experimental research on this subject would, at best, give answers to the ways that independent listener-judges might subjectively react to samples of staged and unstaged talk. However useful such research might be, it still does not answer the question of what language factors account for differences. For this, one cannot do better than to analyze known unstaged and alleged staged speech by the same parties under controlled situations. And this is exactly what the above analyses provide. The results, coupled with the experience that comes from years of listening to conversations, is about the highest level to which this can question can be taken.

THE PERSPECTIVE OF LAW

In order to analyze the way the courts view linguistic analysis of staged versus naturally occurring conversation, it is first necessary to review the evolution of standards permitting expert witness testimony. These standards will then be applied to the Weiss case (2), where such testimony was excluded by the court.

The evolution of expert witness criteria

A review of the judge's ruling in the Weiss case, which excluded linguistic analysis, provides an instructive example of how the judicial system continues to struggle with what should be the simple task of determining when expert testimony may be presented in court. The struggle arises from a basic misunderstanding of the requirements imposed by the 1993 ruling of the United States Supreme Court in Daubert v. Merrell Dow Pharmaceuticals (4).

The ten years of published court rulings since the Daubert decision should have resulted in a resolution of the uncertainty generated by the new standards of admissibility set forth in the Daubert decision. Instead, the judge's ruling to exclude the linguistic analysis in the Weiss case is yet one more example of the confusion that arises when new and unfamiliar issues arrive at the troubling intersection of law and science in American courts.

A resolution of ten years of confusion will assist not only the linguists and lawyers who wish to present the results of linguistic analysis in novel courtroom disputes. A clarification of the confusion resulting from the various misapplications of the Daubert decision should also assist experts in any field who are called upon to apply the methods of their discipline to the resolution of new and unfamiliar courtroom disputes.

Despite the confusion in its application, the law regarding the admissibility of expert testimony in American courts can be simply stated. Until 1993, nearly all American jurisdictions applied a simple standard which was relatively easy to apply. First, the expert providing the testimony must be qualified as an expert in the particular field, and the testimony must be relevant to an issue in dispute in the case. More importantly, for the purpose of this discussion, the scientific theory or technique must be shown to be "generally accepted" by the relevant scientific community. If so, the expert testimony applying the scientific theory or technique was admissible in court.

This "general acceptance" test, originally adopted in the 1923 decision of Frye v. United States (7), came to be widely criticized. Among other deficiencies, the general acceptance test ignored the historical reality that, as science progresses, many theories or techniques, though generally accepted at the time, are ultimately found to be unsound and should be permitted to be challenged in appropriate instances, despite their general acceptance. More significantly, the general acceptance test disallowed the admissibility in court of perfectly reliable methods that were not yet generally accepted in the relevant scientific community simply because they were too new to be well known. The rapid development and refinement of forensic DNA testing from the late 1980's to the present provided prime examples of the deficiency of the general acceptance test set forth in Frye. At its most extreme, prisoners in murder cases were on trial facing the death penalty at the very time that reliable tests were available that could have a key bearing on guilt or innocence, but because of their novelty, could not meet the "general acceptance" test of Frye.

Fortunately, in 1993, tradition gave way to logic, and a new standard emerged. Under the United States Supreme Court's ruling in Daubert v. Merrell Dow Pharmaceuticals (4), to be admissible in court, the scientific theory or technique must be shown to be "reliable"--that is, scientifically sound. Furthermore, under Daubert, the judge must conclude that the expert testimony would "assist the trier of the fact (that is, the jury) to understand the evidence or to determine a fact in issue."

The core of the Daubert decision is the replacement of the "general acceptance" test with a "reliability assessment"--an independent, judicial assessment of the reliability of the offered testimony. As the Court in Daubert explained, a determination of reliability involves an "assessment of whether the reasoning or methodology underlying the testimony is scientifically valid and of whether that reasoning or methodology properly can be applied to the facts in issue."

As guidance in making such a reliability assessment, the Court suggested four factors to assist in the inquiry:

1. Whether the theory or technique has been tested and found to be sound.
2. Whether it has been subjected to peer review and publication.
3. Whether, in respect to a particular technique, there is a high "known or potential rate of error" and whether there are "standards controlling the technique's operation."
4. Whether the theory or technique enjoys "general acceptance" within the "relevant scientific community." (As such, this

fourth factor is a vestige of the old Frye standard.)

The Court emphasized that these are flexible criteria which are meant to be helpful rather than definitive. As an example, in addressing peer review and publication, the Court stated that "Publication .... is not the sine qua non of admissibility," explaining that "in some instances well-grounded but innovative theories will not have been published."

The Court in Daubert also pointed out that the four criteria are not an exclusive or invariable list. Because of the wide variety in types of expert testimony, factors that are helpful in determining the reliability of one form of expert testimony (DNA testing, for example) may be inapplicable to a reliability assessment in another field (such as psychiatric diagnosis).

Upon the issuance of the Daubert decision in 1993, the four-part reliability assessment became immediately applicable to all federal court proceedings involving scientific testimony. In the years since 1993, the Daubert standard has also been adopted by an increasing number of state courts.

Despite the apparent clarity of the Daubert criteria, much confusion ensued. One key example of the confusion was the immediate controversy regarding the scope of the ruling. The debate involved whether the Daubert standard applied only to testimony regarding scientific testing and analysis, or whether it applied to all forms of expert testimony, including testimony by those whose expertise comes from training and experience, such as engineers, automobile mechanics, and therapists.

Unfortunately, because of the slow and cumbersome operation of the legal system, six years of uncertainly followed before the issue reached the United States Supreme Court for clarification. Finally, in 1999, in Kumho Tire Company v. Carmichael (8), the Supreme Court addressed the subject again, resolving much of the ambiguity left unaddressed in Daubert. The Court in Kumho Tire ruled that the reliability assessment set forth in Daubert applies to all forms of expert testimony, if challenged, not just to the narrow category of scientific tests. Additionally, the Court emphasized that the four-part Daubert test "was meant to be helpful, not definitive."

Most significantly, the Court's decision in Kumho Tire resolved many of the issues that bear on the admissibility of applying established scientific methodology to the resolution of novel questions. The Court states that "It might not be surprising in a particular case, for example, that a claim made by a scientific witness has never been the subject of peer review, for that particular application at issue may never previously have interested any scientist."

Such novelty of application may cause the proposed testimony to fail one or more of the particular Daubert criteria (peer review and publication, for example) but does not eliminate the particular testimony from admissibility in court. As the Supreme Court explained, "the factors identified in Daubert may or may not be pertinent in assessing reliability, depending on the nature of the issue, the expert's particular expertise, and the subject of his testimony . . . Too much depends upon the particular circumstances of the particular case at issue."

The reliability assessment is easiest to apply to the clear-cut examples at the extremes of the spectrum. Despite the continued litigation about the details, all courts now appear to accept the underlying reliability of DNA testing, for example, while none would accept the testimony of an astrologer or a palmist.

Greater difficulty is encountered when the subject of expert testimony involves scientific studies of various aspects of human nature. Consequently,

courts frequently struggle with Daubert challenges to such disputed matters as diagnoses of multiple personality disorder or sexual abuse accommodation syndrome, or testimony regarding the reliability of eyewitness testimony or repressed memory.

Unfortunately, the very flexibility of the Daubert criteria gives rise to increased opportunity for misunderstanding and abuse. Judges are required to weigh the various flexible criteria, without the use of scales, and simply exercise their discretion and pronounce whether, in their judgment, the expert testimony is sufficiently reliable to be presented to the jury. And judges all have their biases, consciously or unconsciously, towards particular forms of evidence and toward the parties proposing to introduce the evidence in court.

The very flexibility of Daubert criteria allows judicial biases to come into play and permit judges to ignore or override the purposes that Daubert criteria were designed to serve. Aside from the interplay of conscious and unconscious bias, the flexibility of Daubert criteria can also give rise to well-intentioned but unfortunate misunderstandings and erroneous conclusions in specific cases.

Application of Daubert to the Weiss case

A review of the judge's ruling excluding the linguistic analysis in the Weiss case provides examples of the concerns found in the evolution of the Daubert criteria. In this case, the judge surprisingly concluded that the linguistic analysis about whether or not the conversations were staged was based on "uncharted waters with little indicia of reliability." The judge justified this conclusion by applying Daubert criteria, but in doing so, she applied these criteria in the narrowest and most inflexible manner and achieved a result that appears to be contrary to the basic notions of a reliability assessment.

For unexplained reasons the judge applied the Daubert criteria in a rearranged order. First, the judge addressed the criterion that includes the traditional "general acceptance" test, the vestige from the Frye test. The judge concluded, without setting forth any explicit reasoning, that the technique involved in the linguistic analysis in this case "is not generally accepted in the scientific community." This unexplained observation appears to be based on the judge's narrow definition of the term, "technique," which is equated with applying linguistic analysis to a specific determination of whether or not the conversations appear to be staged.

When so narrowly defined, however, no expert testimony applying established techniques to new facts will ever pass the Daubert challenge. The Daubert criteria were not intended to be applied so restrictively. As the United States Supreme Court clearly stated, Daubert challenges are not intended to exclude a technique simply because it is the first time the technique has been applied to a particular type of issue. In this particular case, the specific question being asked is a new one, but the techniques applied to help answer the question are long-established and accepted practices of linguistics in general and forensic linguistics in particular.

What is crucial to a reliability assessment is whether the method, that is, the technique that is being applied, is reliable. The technique involved in this case is a linguistic analysis of the various discourse and articulation factors contained in both natural speech and in the questioned conversations. The ability of such analysis to identify linguistic characteristics of speech and of the circumstances surrounding the speech has long been established. The reliability of this method in identifying linguistic characteristics is not subject to serious challenge.

With the proper focus on the method in question, even an inflexible application of Daubert criteria results in admissibility of the evidence. If the question is properly phrased, "Is the method generally accepted within the relevant scientific community?" the answer is an unequivocal "Yes."

With regard to the second Daubert criterion, peer review and publication, the same result is obtained. The judge in the Weiss case rejected the testimony on the grounds that there has not been peer review and publication of the narrow question of analyzing staged speech. There has not been peer review and publication of studies of this narrow question simply because the very question had never been posed. But the technique involved, the application of linguistic analysis to natural speech and to questioned conversations, has been repeatedly subjected to peer review and publication. And as Daubert and Kumho Tire make clear, peer review and publication are not required, and indeed cannot be required, for scientific studies that have never previously been conducted.

The third Daubert criterion that the judge addressed in this case, whether the theory or technique has been tested and been found to be sound, once again hinges on how broadly or narrowly the criterion is applied. The application of linguistic analysis to naturally occurring human speech in virtually all contexts and genres has been tested frequently and found to be reliable. The application of linguistic analysis to the much more narrowly-defined issue of staged conversation has not been tested, however, and would be far too difficult, if not impossible, to test. Staged speech is contrived and it occurs unnaturally. Any possible sample of staged speech data would be subject to variables relative to the skills of a speaker to emulate naturalness, which is the primary reason why comparing known speech samples of natural speech with samples of allegedly staged speech is so important.

Still, the ability to perform such testing may be an important issue in determining the admissibility of applying established techniques to novel issues. As the Court stated in Daubert, "The criterion of the scientific status of a theory is its falsifiability, or refutability, or testability." Although the United States Supreme Court in Daubert specified that all criteria do not have to be met for proposed expert testimony to be admissible, the judge in Weiss is certainly correct in considering testability to be a factor.

Finally, the judge in Weiss applied the fourth Daubert criterion, and ruled against the admissibility of the linguistic analysis because there was no known or potential rate of error that was identified for the linguistic analysis in this case. On one level, with regard to that specific linguistic analysis, the absence of a known or potential rate of error should not have been an issue because the specific data compiled in the linguistic analysis was presented in an observable fashion, verifiable by all parties. The number of occurrences of overlapping speech, interruptions, topic recycling, contractions, pause fillers, intonation patterns, etc., are all reviewable for accuracy by anyone involved in the proceeding. Error rates are simply not an issue when the data can be independently confirmed by anyone who wishes to do so.

On another level, however, error rates may matter in a reliability assessment if the issue of error rate is applied to the ultimate question of whether such characteristics (overlap, interruptions, contractions, etc.) are characteristic of natural speech at all. To some extent the answer to this question may be intuitively obvious, but not precisely measurable. It is for this reason that the factors set forth in the Daubert reliability assessment are flexible ones. As the United States Supreme Court acknowledged in Kumho Tire, each Daubert factor may not be applicable in every case: "The factors identified in Daubert may or may not be pertinent in assessing reliability, depending on

the nature of the issue, the expert's particular expertise, and the subject of his testimony."

In rejecting the linguistic analysis in Weiss, the judge expressed concern that the staging of a conversation could be done so skillfully that the speakers could adopt the characteristics of natural speech and avoid detection. The linguistic analysis would then appear to confirm that the speech was natural only because the speakers, in their artifice, had skillfully adopted the characteristics of natural speech. Aside from the absence of any indication that the speakers acquired a sophisticated background in linguistics, and had the theatrical ability to implement it, the judge's conclusion is nevertheless erroneous. As a matter of law, the theoretical possibility of deception does not preclude the introduction into evidence of expert testimony on a particular subject. It is widely acknowledged, for example, that crime scene evidence can be fraudulently placed at a scene. The possibility of planting a blood stain, however, does not preclude the introduction of testimony about DNA testing on the stain. If the possibility that evidence can be fictitiously manufactured or planted would preclude admissibility, then no DNA testing, fingerprint comparison, chemical analysis, or psychiatric or psychological testimony could ever be presented in court.

Finally, the judge rejected the admissibility of the forensic analysis on the grounds that it would not "assist the trier of fact" in understanding the evidence or in determining a fact in issue. The judge explained her reasoning by stating that "An expert witness on the issue of whether these tapes were 'staged' or 'not staged' does not add anything helpful to the jury's own consideration of the direct evidence and the credibility of the witnesses with direct knowledge of the taped conversation and the creation of these exhibits..."

In light of the information gleaned from a linguistic analysis in this case, this conclusion by the judge is particularly difficult to comprehend. The comparison of overlapping speech, interruption rates, topic recycling, contractions, intonation, etc., adds so much to the determination of whether the conversations were natural or staged that it appears that the judge's real concern may not have been that the analysis would not assist the trier of the fact but that it would instead assist the trier of the fact too much. Although it is impossible to state the existence of judicial bias with any degree of certainty, experienced criminal defense lawyers have long ago noted how difficult it is to get a judge to allow testimony on behalf of a client that a judge, rightly or wrongly, perceives as a rogue.

Taken as a whole, the judge's ruling in Weiss appears to be a narrowly defined, inflexible application of Daubert criteria, resulting in the mistaken exclusion of sound and established linguistic analysis which could have helped to resolve a new and interesting issue in court. Fortunately, both expert witnesses and lawyers can learn from the unsuccessful efforts to introduce scientific analysis into the courtroom. While emphasizing the need to apply Daubert criteria broadly and flexibly, lawyers and experts can simultaneously prepare to meet the narrowest possible interpretations of Daubert criteria. And both scientist and lawyer can take heart in the improvements in Daubert rulings as courts become more proficient and more accurate in making reliability assessments. Early rulings excluding DNA results, for example, now seem so clearly mistaken that it is unlikely that DNA evidence will ever again be excluded from courtrooms on the grounds of being unreliable. As judges become more adept at making reliability assessments, and as correct rulings begin to outweigh incorrect ones and gain in precedential value, reliability assessments

will be more accurately made in the more difficult cases, where aspects of human behavior that are poorly understood and difficult to measure are at issue.

Alternatives to Linguistic Testimony

In the meantime, lawyers are not at a total loss where such matters as the linguistic analysis in the Weiss case are excluded from court. Because of the logical basis for much of the analysis, the lawyer can present much of the analysis to the jury without calling the expert to the stand. For example, after studying the linguistic analysis with the expert, the lawyer can present much of it, albeit not as effectively, without the testimony of the expert.

In opening statements to the jury, for example, the lawyer is permitted to explain what the evidence in the case is expected to be. The lawyer can simply explain that the evidence will involve taped statements that are difficult to understand: that is, taped statements that contain language that is filled with overlaps, interruptions, confusion, arguments, complaints, sarcasm. As the tapes are introduced as evidence, the lawyer can then ask the person introducing the tapes on the witness stand to count such matters as the number of interruptions and overlaps. Finally, in closing argument, where the lawyer is permitted to argue all reasonable and logical inferences from the evidence, the lawyer can take the benefit of the linguist's analysis and explain to the jury as emphatically as possible why the various characteristics support the conclusion that the questioned speech is natural rather than staged. The logic behind the analysis (the very reason that the linguist's analysis should have been admitted in the first place) permits the lawyer to make many of the same points to the jury even in the absence of the linguist as a witness.

The reliability assessment mandated by the Daubert decision has forced both lawyers and judges to address, for the first time, the underlying validity of the expert testimony that is presented in court. The initial steps have been clumsy at times, but the judicial system as a whole will continue to benefit as lawyers and judges overcome the initial clumsiness and become increasingly adept at addressing the merger of science and law in the resolution of judicial disputes.

References

1. Shuy, R.W. What we do with English when we take notes: evidence from a civil suit. Studia Anglica Posnaniensia 1998;33:353-63.
2. United States v. Sholam Weiss, No. 98-99-Crim-Orl-19A (M.D. Fla. September 2, 1999).
3. State of Tennessee v. Byron Looper, No. 5346 (Crim.Ct. Cumberland Co.) (tried August, 2000, appeal pending.)
4. Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993).
5. Tiersma, Peter. 1999. Personal correspondence.
6. Solan, L. Linguistic experts as semantic tour guides. Forensic Linguistics 1998;V2: 87-106.
7. Frye v. United States, 293 F. 1013 (1923).
8. Kumho Tire v. Carmichael, 526 U.S. 137 (1999).