In: Language and Linguistics Emerging Trends
Editor: Cynthia R. Dreyer, pp.

ISBN 978-1-60456-893-6

Chapter 7

THE NEED FOR EXPLICIT INFERENTIAL METHODS IN LINGUISTICS

Kent Johnson
Department of Logic and Philosophy of Science
University of California, Irvine

ABSTRACT

Like every discipline, theorizing in linguistics crucially involves drawing (often tentative) conclusions about the relation between the available evidence and the theories under consideration. At many points in their research, linguists must decide whether the available evidence (theoretical, empirical, etc.) supports a given theory well enough to accept it, possibly over various competing theories. Surprisingly, however, virtually no attention has been paid to this aspect of linguistic theorizing, which contrasts with the standard practices of the social, behavioral and psychological sciences. In the latter fields, whole subareas – including dedicated faculty positions, journals, societies, etc. – are devoted to the quantitative, frequently statistical, details of drawing inferences from data. But in linguistics, theoretical inferences typically have the form of informal, verbal, holistic judgments made by professional linguists. I begin by characterizing the issue of linguistic inferences, showing why the current methods are cause for concern. The psychological literature on professional judgment strongly suggests that professional judgments aren't as accurate as the experts believe, and that they are inferior to more explicit methods. In fact, linguistic inferences are particularly (although not uniquely) suspect in this respect. I then consider several responses linguists have made to the criticism of their reliance on expert judgment. These replies, I argue, do little to mitigate the concerns about how evidence is compiled and assessed in linguistics.

1. INTRODUCTION

Let's play a game. Consider the two sequences in (1) and (2). One of them represents the beginnings of a sequence that resulted from flipping an unbiased coin 200 times. The other one, the "fake", was not generated by flipping a coin. If you correctly identify the fake, I will give you $50, but I will charge you $50 if you are wrong. If you choose not to answer, I will give you $5 for your modesty. (The reader can also play the game with the two sequences of length 30 below; cf. footnote 2.) What shall you do?

(1) T H T H T H T T T T H H T H T T T H T H H H T T T H H T T H …
(2) H T T H H T T T H H T H H T H T T H H H T T H T T H H T T H …

Simply staring at (1) and (2) is unlikely to produce any judgments that can be confidently held. Moreover, all size 200 sequences are equally likely to be produced by a fair coin, hence from this perspective (1) and (2) are equiprobable. To make the game more interesting, I also mention that I constructed the fake sequence with an eye towards how often Heads would be in the lead[1].

Aha! Now you have something to work with. You recode the two sequences and discover that Heads is never in the lead in (1), but is in the lead about half the time in (2). While it is clearly *possible* that Heads would never have the lead in 200 flips of a fair coin, it is also clearly much less likely than the other outcome. So you now seem to be in a position to infer that (1) is the fake sequence and (2) is genuine. But how strong is this evidence? Is the case for (1)'s being fake so strong that you should be quite comfortable drawing this conclusion? Or is the evidence only moderately strong, so that you should be cautious about drawing a conclusion? Or maybe it is quite weak, so that you should still make no guess and just take the $5?

At this point, you reason that the present scenario is restricted to only two extremely opposed options – share the lead equally versus never having the lead even once. If there were more options, your confidence might be eroded, perhaps significantly. But in the present case, you stand by the logic that a fair coin would put Heads in the lead about as much as Tails is in the lead, and so you officially tender your guess.

Walking away from this game $50 poorer, you wonder what happened. Was it simply a statistical fluke, and the fair coin never let Heads take the lead even once? Or could this outcome have been due to something else? It was: your assessment of the evidence was incorrect. Although it doesn't suit our intuitions about randomness and probability, the most likely outcomes are where either Heads or Tails is in the lead the entire time; the least likely outcome occurs when the lead is shared equally. When flipping a fair coin 200 times, there is more than an 11% chance that one side will have the lead the entire time, compared to a 1% chance that the lead will be shared equally. Similarly, there is almost a 25% chance that one side will be in the lead at least 97% of the time, but there is less than a 4% chance that the lead is shared equally within the corresponding margin of 3% (cf. Feller 77ff. for proofs and discussion).[2]

The relatively simple scenario just described illustrates an important phenomenon: even in this little game, it is hard to it is hard to come up with a reliable assessment of the evidence just by eyeballing the evidence. Human minds, even those of expert mathematicians and

[1] More precisely, Heads is in the lead after the *k*th flip if either (a) there are more Heads than Tails in the *k* flips, or (b) there are exactly *k*/2 Heads (and hence *k*/2 Tails), and Heads was in the lead after the *k*–1th flip.

[2] Incidentally, the actual sequences in (1) and (2) are both only of size 30, with ones in the lead not even once in (1), and 18 times in (2). Even in these truncated sequences the probabilities are telling: there's a 14% chance that Heads will be in the lead the entire time, and a 4% chance that one side will be in the lead exactly 18 times.

statisticians, are not built to make accurate[3] on-the-fly assessments regarding how a set of evidence bears on the relevant hypotheses. This basic fact has been studied intensively for decades in the decision sciences community. The purpose of this chapter is to stress the importance to linguistics of this well-known limitation on human reasoning. Much of mainstream linguistic theorizing is built upon a practice of informally aggregating evidence in situations where we have little reason for confidence regarding the accuracy of such "expert judgments". Indeed, although it is rarely acknowledged in the linguistics literature, I will argue that the particular inferential situation that the linguist finds herself in is an extremely difficult one, and far more problematic than many other areas that routinely appeal to informal expert judgments.

2.THE INFERENTIAL SITUATION IN CONTEMPORARY LINGUISTICS

To get things started, let's examine a fragment of actual linguistic theorizing. The example below is representative of much of the methods of mainstream linguistic theorizing. Since the purpose of the example is only to illustrate these widely used *methods*, I will not try to exhaustively characterize the relevant literature. For present purposes, that would only obfuscate matters. (Indeed, the precise details are unimportant enough that readers familiar with linguistic methods may wish to skim the example, and go right to §3.)

In a series of papers, Norbert Hornstein (1998, 1999, 2000, 2003) argues that the phenomenon of linguistic control can be accounted for simply by allowing movement into theta positions. E.g., the relevant syntax of (3a) does not have the traditional form in (3b), where PRO is a distinct lexical item controlled by *Sue*. Instead, the proper form is in (3c), where *Sue* has moved from the lower subject position to the higher one. (Following Hornstein, I treat movement as a combination of the Minimalist operations of Copy and Merge.)

(3)

a. Sue wants to win;
b. Sue$_i$ wants [PRO$_i$ to win];
c. Sue wants ~~Sue~~ to win.

More generally, Hornstein holds that linguistic theories do not require the null pronominal element PRO or its associated control module that determines the referent of an occurrence of PRO. We don't need these things, Hornstein argues, because the phenomena that initially motivate positing them can be accounted for by appealing to independently motivated components of the grammar. Movement (aka Copy and Merge), Hornstein assumes, is a prevalent feature of grammar. If all the relevant facts can be accounted for without positing PRO, then ceteris paribus, linguistic theories should favor the simpler theory and reject the

[3] I use the term *accurate* as a catchall phrase to refer to the various qualities that make for good judgments. These include most prominently validity (the tendency for judgments to center around the normatively correct answer) and reliability (the tendency for judgments to cluster relatively tightly together).

employment of PRO. In the development of this theory, Hornstein also notes several advantages. Here are two representative examples.

Exhibit #1: de se interpretations. Consider (4a). Notice that it can only have the meaning given in (4b), and cannot have the meaning of (4c):

(4)

a. Only Bush remembers giving the ‘03 State of the Union Address.
b. Only Bush remembers himself giving the ’03 SOTU.
c. Only Bush remembers that he gave the ’03 SOTU.

(4c) is false, since many persons remember that Bush (and only Bush) gave the address. However, since Bush (and only Bush) gave the address, he (and only he) *even could* remember giving it; thus, the truth of (4a) depends only on what Bush (and only Bush) remembers.

What accounts for this mandatory *de se* reading? As Hornstein notes, if control is really a form of raising, the phenomenon in (4) is handily explained. After all, in such a case, the relevant syntax of (4a) would be as in (5a):

(5)

a. [$_{\text{IP}}$ Only Bush [$_{\text{VP}}$ ~~Only Bush~~ [remembers [$_{\text{IP}}$ ~~Only Bush~~ [$_{\text{VP}}$ giving the ‘03 SOTU]]]]]
b. Only Bush λx.[x remembers x giving the ‘03 SOTU]

If the traces (i.e., results of Copy and Delete) in (5a) function as free variables bound by *Only Bush*, then the syntax easily supports the reflexive predicate reading in (5b), which necessitates a *de se* reading.

Exhibit #2: Similarity to raising. As another piece of evidence, Hornstein notes some similarities between control and raising patterns. In the raising patterns in (6), we see that the raising must Merge a copy of the deleted DP in the expression, and that the raised expression must Merge in a way that respects subjacency restrictions:

(6)

a. John seemed ~~John~~ to be happy.
b. *It seemed ~~John~~ to be happy.
c. *John thought that it seemed ~~John~~ to be happy.

An identical pattern is seen with control structures:

(7)

a. Cathay was expected to be happy.
b. *It was expected to be happy.
c. * Cathay thought that it was expected to be happy.

In short, the similarity of behaviors of control and raising structures further suggests that they may be from a common stock. Since there are, Hornstein notes, independent reasons for positing raising mechanisms, the similarities just noted provide further evidence that control is simply raising.

Unsurprisingly, Hornstein's proposal has not gone unnoticed (e.g., Brody 1999, Culicover and Jackendoff 2001, Landau 2000, 2003, Manzini and Roussou 2000). The following two exhibits present prima facie evidence against the view that control is a type of movement.

Exhibit #3: Partial Control. The first problem comes from Landau (2003), who argues that Hornstein's theory has problems accounting for "partial control". For example, consider (8):

(8) The chair of the department wanted to meet on Tuesday afternoon.

As Landau notes, the most natural interpretation of (8) is that the chair of the department wanted some group that contains the chair and at least one other person to meet on Tuesday afternoon. That is, the controlling DP only partially determines the subject of the lower clause. But it is very hard to see how the remainder of a raising process, ~~the chair of the dept.~~, could support this 'group' reading. Moreover, the lower clause can contain predicates that demand that the subject be a non-singleton group: e.g., *Susan enjoyed getting together on weekends, Steve wondered whether helping one another would be productive in the long run*. Worse yet, classical raising constructions do not seem to support these group readings:

(9) *The chair seemed to meet on Tuesday afternoon.

(9) cannot mean that some group containing the chair seemed to meet on Tuesday. In short, Landau's evidence regarding partial control suggests that there must be distinct elements in the relevant positions of raising and control constructions.

Exhibit #4: Control and NP movement. A second bit of evidence against Hornstein's theory comes from Brody (1999, 218 – 19). If control and raising were a unified phenomenon, we would expect them to exhibit identical behavior; but they don't. If raising were simply control, then just as the control structure (10a) can express that John attempted to make it the case that John leave, so too, there should be a similar possibility of raising, allowing (10b) to express that John believed himself to have left, which is not the case:

(10)

a. John attempted to leave.
b. *John believed to have left.

In the other direction, we would also expect that just as the passive raising in (11a) is acceptable, having the meaning that it was believed that John had left, so too should (11b) be acceptable, with the meaning that some agent(s) attempted to make John leave.

(11)

a. John was believed to have left.
b. *John was attempted to leave.

These disparities between raising and control structures further suggests that the latter phenomenon is not an instance of the former.

Obviously there is much more to be said about the Hornstein's view. In addition to a great many more sources of evidence – both supporting and undermining Hornstein's view – there are also replies and reanalyses of the evidence, and replies and reanalyses of the replies and reanalyses, etc. However, it is enough for present purposes to consider the simple situation characterized above.

In this mini example, we have four considerations: Hornstein's theory does well at accounting for (i) some similarities between raising and control, and (ii) some phenomena concerning *de se* interpretations. However, it has difficulties with (iii) partial control and with (iv) some issues concerning NP movement. Let us suppose (fictitiously, of course) that this is all the available data regarding Hornstein's theory. How good is the theory? Does it merit provisional acceptance? Is it promising enough that a linguist working in a related area should explore how the theory interacts with hers? All else being equal, is it better than a theory that does well with (iii) and (iv), but not so well with (i) and (ii)? If so, by how much? These are the sorts of questions that linguists regularly face and address in their research, whether tacitly or explicitly. Crucially, the dominant trend in linguistics is for the linguist to compile and assess the available evidence using informal, holistic, purely verbal methods. That is, linguists examine bodies of evidence – typically much larger and more complex than just (i) – (iv) above – and arrive at a judgment as to what it says about the competing theories. In particular, these judgments regarding the theory's relation to the evidence are made without the use of any "external" aids, such as statistical rules or models, or other mathematical methods.

The introduction to this paper presented a particular instance where our intuitions were much less accurate than a mathematical assessment of the evidence. This was so despite the fact that the evidence and theories (i.e., which sequence was the fake) were vastly simpler than what linguists typically work with. In the present section, we saw that linguists typically assess how the evidence bears on their theories solely by means of their expert judgments. The task of §3 is to argue that, in terms of their accuracy, expert judgments in linguistics are probably disturbingly similar to our faulty judgments seen in §1.

3. The Need for Explicit Inferential Methods

The previous two sections raise a general issue: linguists typically assess the theory/evidence relationships with informal, holistic, verbal judgments, but this method may be unreliable. Importantly, reliance on informal judgments is widespread in linguistics. Johnson 2007 notes 40-odd articles and books in the linguistics literature where every inference, assessment, etc. has the same general form as that depicted in §2. But there is nothing special about those texts: readers of the major linguistics journals (e.g. *Linguistic Inquiry, Natural Language and Linguistic Theory*, etc.) know that this list could be increased by an order of magnitude with little difficulty. Rather than being a few notable exceptions, reliance on the informal, verbal, holistic judgments of professional linguists is the dominant trend in linguistics.

But what about the example in §1? That was just one particular case where our intuitions about the evidence are faulty. Could this be an isolated case? Maybe our intuitions about evidential strength – or at least the intuitions of those who are experts about the relevant topics – are generally quite good. If so, then there may not be much of a methodological problem after all. The burden of this section is to show that the inaccuracy illustrated in §1 is prevalent even in expert judgment.

To begin, some terminology. The general phenomenon of interest is often called *expert judgment*. It concerns the abilities of professionals and other experts to aggregate diverse sources of evidence (from within their field of expertise) and arrive at high quality judgments, predictions, decisions, etc. By definition, these judgments are informal, verbal, largely holistic assessments of a body of evidence. Crucially, such judgments are produced without the aid of a statistical or mathematical analysis of the evidence. This type of ratiocination should not be conflated with human reasoning in general, which may exhibit flaws not found in expert judgments. Indeed, it is natural to suppose that expert judgments should be considerably better than those produced by typical human subjects. After all, experts are *experts*: they have spent many years studying the topic(s) that they are supplying judgments about. Moreover, reliance on expert judgment is a common and well-accepted cultural practice.

Although one would expect that expert judgments suffer from few if any of the foibles that ordinary human reasoning does, unfortunately this is not so. This can be illustrated with a couple famous examples. (I canvass several examples in order to stress the importance of acknowledging the negative evidence regarding expert judgment; e.g. Arkes 2003. Hopefully, the concreteness of the examples will emphasize just how different experts' judgmental capacities are from how they are commonly perceived.)

Many readers of this chapter are familiar with the task of graduate admissions: the committee needs to decide which applicants to admit, which to recommend for fellowships, assistantships, etc. Many readers may even consider themselves quite skilled at detecting which applicants have the greatest potential to become fellow professionals in their field. But such a task involves aggregating diverse sources of evidence, typically of much less complexity than the diversity of linguistic evidence. Experts and non-experts are typically quite bad at such aggregation tasks. Consider, for example, the comparison of two candidates from comparable undergraduate schools. Candidate A has a GPA of 3.3 and a GRE of 750; candidate B has a GPA of 3.7 and a GRE of 680. Which candidate is more desirable? Why? By how much? Even in this familiar comparison task, it is unclear how the candidates should be ranked. Regarding this example, Robyn Dawes writes that

> most judges would agree that these indicators of aptitude and previous accomplishment should be combined in some compensatory fashion, but the question is how to compensate. Many judges attempting this feat have little knowledge of the distributional characteristics of GREs and GPAs, and most have no knowledge of studies indicating their validity as predictors of graduate success. Moreover, these numbers are inherently incomparable without such knowledge, GREs running from 500 to 800 for viable applicants, and GPAs from 3.0 to 4.0. Is it any wonder that a statistical weighting scheme does better than a human judge in these circumstances?" (Dawes 1979, 574)

In another paper (1971), Dawes attempted to estimate the accuracy of the expert graduate admissions judgments about each applicant's potential for success. The correlation was essentially zero: $r^2 = .0361$. Indeed, I posed these questions about combining GRE scores and GPAs to a senior colleague whose career has centered around the decision sciences. His response to them was straightforward: "I don't know".

Notice incidentally that this comparison task involving aggregating evidence is the simplest possible: two candidates measured on two differently scaled variables. Thus, this task has a structure that is similar to, but simpler than, the problem the previous section ended with. In contrast, the corresponding task in linguistics is massively harder: there are typically many disparate sources of evidence which are hardly ever precisely located on an (approximately) rational scale. Instead, the evidence is typically verbal, and often vague in terms of its exact characterization. Moreover, there are often more than two candidates (i.e., candidate theories) to be evaluated, and often some of these sources are inapplicable to some of the candidates.

The second example is a famous case where Tversky and Kahneman (1971) tested the expert judgments of psychologists with a problem that was at the heart of their expertise. In a questionnaire given to members of the Mathematical Psychology Group and the American Psychological Association, they asked:

> Suppose you have run an experiment on 20 subjects, and have obtained a significant result which confirms your theory ($z = 2.23$, $p < .05$, two-tailed). You now have cause to run an additional group of 10 subjects. What do you think the probability is that the results will be significant, by a one-tailed test, separately for this group? (Tversky and Kahneman 1971, 105).

Here again, we would expect these expert judgments to be highly reliable. In fact, they were not. The median response of these experts was .85. However, Tversky and Kahneman argue that a much more reasonable estimate is .48 or lower, a difference of 37 percent.

It is noteworthy that the respondents were asked to evaluate a fairly simple situation. All that is involved is a basic form of null hypothesis testing with a random sample drawn from a normally distributed population. Any introductory statistics course will cover the concepts employed in this question. Thus, the low quality judgments these experts made – about a question that could've easily arisen in a lower-level undergraduate course that they teach – is all the more striking. Moreover, the respondents were trained academic psychologists, and statistics and probability are their main methodological tools. Thus, if anything we should expect them to be particularly good at making judgments under conditions of uncertainty (cf. Nisbett 1993).

To put it mildly, these classical results are not alone. Over the decades, there has developed a mountainous literature regarding the nature of professional judgment. A consistent theme in this literature is the drawing of "pessimistic conclusions" (Arkes, Dawes and Christensen 1986, 93). For example, Camerer and Johnson conclude their paper with:

> Our review produces a consistent, if depressing, picture of expert decision-makers. They are successful at generating hypotheses and inducing complex decision rules. The result is a more efficient search of the available information directed by goals and aided by the experts'

> superior store of knowledge. Unfortunately, their knowledge and rules have little impact on experts' performance. Sometimes experts are more accurate than novices (though not always), but they are rarely better than simple statistical models. (Camerer and Johnson 1991, 211)

In particular, experts are quite bad at aggregating various sources of evidence into normatively appropriate judgments (or estimations and the like). In fact, a great deal of research has shown that even very crude statistical models outperform expert judgments (Meehl 1954, 1986, Dawes 1971, 1979, Dawes and Corrigan 1974, Lovie and Lovie 1986, Larrick 2004, Bishop and Trout 2005). As Meehl notes in a retrospective essay:

> When you are pushing 90 [the current count is about double this –KJ] investigations, predicting everything from the outcome of football games to the diagnosis of liver disease and when you can hardly come up with a half dozen studies showing even a weak tendency in favor of the [sc. expert over the statistical model], it is time to draw a practical conclusion (Meehl 1986, 374).

(Meehl goes on to argue that the extremely few cases where the expert outperformed the statistical model may well be due to statistical sampling error.) As part of this research, investigators have found many different causes behind the faulty judgmental capacities of experts. Here are just four well-known examples.

First, experts and non-experts alike tend to give excess weight to novel information (Shafir and LeBoeuf 2004). E.g., when evaluating college applicants, people tend to assign more weight to a grade-point average that appears after the rest of the dossier has been considered than when the GPA is included in the dossier.

Second, harder problems tend to increase experts' degree of overconfidence in judgments E.g., in one famous study, subjects tended to express more overconfidence regarding harder questions (for which the subjects' error rate was higher), and less overconfidence for the easier questions (Lichtenstein et al. 1982, Harvey 1997, Griffin and Brenner 2004).

Third, experts with little training in normative judgmental models (e.g. statistical models) tend to show even more miscalibration with the normatively appropriate judgments than those with greater familiarity and training (Griffin and Brenner 2004).

Finally, in their judgments, experts tend to reflect systematic biases in favor of their own theories, treating the evidence for their own theories (and the evidence against rival theories) as overall stronger than it actually is (e.g., Nickerson 1998, Harvey 1997, Larrick 2004, Griffin and Brenner 2004, Camerer and Johnson 1991).

At this point, an obvious question arises. If things really are this bad, why is expert judgment still used in these tasks? Surely most experts and organizations would not want to expend large amounts of time and money carefully collecting evidence, only to assemble it in a way that vastly undermines its probative quality.

This question too has been the subject of immense research. While there are many things to say about this issue, one stands out: experts are overconfident, often wildly so, about their abilities.[4] E.g., Faust and Ziskin write of "the immense gap between experts' claims about

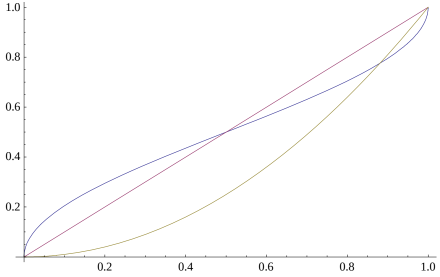
[4] Notice the difference between the confirmation bias mentioned at the end of the previous paragraph and the judgmental overconfidence under discussion. This difference can be seen by comparing psychometric curves relating expert judgment regarding the probability of some statement on the *x*-axis with the normatively appropriate probability on the *y*-axis (e.g. Harvey 1997). The diagonal line shows perfect calibration between the judged and correct probability. The convex line reflects a confirmation bias, where the judged probabilities

their judgmental powers and the scientific findings” (Faust and Ziskin 1988, 35). Overconfidence about judgmental powers even appears when the evidence is highly precise and is embedded within a powerful mathematical theory. For instance, Fischoff and Henrion’s well-known (1986) study of confidence levels of physicists’ estimations of various physical magnitudes (e.g., the velocity of light or the rest mass of the electron) revealed a constant trend of overconfidence. In short, when it comes to the aggregation of evidence in their field, experts are typically unskilled and unaware that they are so (Kruger and Dunning 1999). Moreover, this overconfidence regarding expert judgment appears to be solidly entrenched in contemporary culture. E.g., even organizations as prestigious as the National Science Foundation have elected not to heed the cautionary tale from psychology about expert judgment (Arkes 2003).

The relevance to linguistics of the literature described above is clear. There is no reason to believe that the expert judgments in linguistics are of any higher quality than the expert judgments reviewed above.[5] It is not unreasonable to suspect that linguistic judgements are vulnerable to the four examples of limitations on expert judgment. It would be easy for a linguist to (tacitly) treat some newly discovered data from an exotic language as having more probative quality than it perhaps should, simply because it is “new to the scene”. It would also be easy for a linguist to have more confidence than is warranted in the fundamental architecture of her linguistic theory: knowing whether, e.g., Minimalist syntax is on the right right track – and in what respects – is clearly a very hard question. Moreover, the standard graduate-level training in linguistics does not include coursework (e.g., probability and statistics) which focuses on methods for objectively collecting, organizing, and analyzing the evidence so as to manage the uncertainty present in the data and assess its bearing on the theories at hand. Finally, unless linguists are unlike other experts, it is reasonable to suppose that, when “telling a good story” about their theory, linguists will at times selectively recruit and overweight the evidence in favor of the theory, and underweight the evidence against it. (And of course these features could easily reappear in criticisms of a given theory, and in criticisms of the criticisms, and so forth.)

In fact, linguistic judgments are in a particularly precarious position, since these judgments are not honed with the aid of feedback. It is not, for instance, part of the standard graduate-level linguistics education for students to be given evidence-theory pairs, where the

(regarding one’s favored theory) is systematically higher than the true probability. The sigmoidal line displays judgmental overconfidence, where improbable statements are judged to be more improbable than they actually are, and probable statements are judged to be more probable than they are. (Continued)



[5] Notice also that there is a skeptical facet to this difficulty. Although the evidence reviewed above give much reason for pessimism, in fact it is utterly unknown how well expert linguistic judgments typically perform. Since there is relatively little consensus regarding nontrivial issues in linguistics, it would be hard (although perhaps not impossible) to find issues which can be treated as settled, and examine the relevant inferences drawn about them in the past literature, to gauge the nature of the judgmental mechanisms.

theories are at or near the level of detail as Hornstein's and are known to be correct or incorrect, and told to assess how well the evidence supports the theory. If that were possible, then since the actual status of the theories in this "training set" is known, the feedback would allow for the budding experts to calibrate their judgments. At the same time, mere feedback is not enough. For example, Camerer and Johnson write that:

> One of the main lessons of decision research is that feedback is crucial for learning. Inaccurate [rules] may persist because experts who get slow, infrequent, or unclear feedback will not learn that their rules are wrong. When feedback must be sought, inaccurate rules may persist because people tend to search instinctively for evidence that will confirm prior theories….Even when feedback is naturally provided, rather than sought, confirming evidence is more retrievable or "available" than disconfirming evidence…. The disproportionate search and recall of confirming instances will sustain experts' faith in inaccurate [rules]. Even when evidence does disconfirm a particular rule, we suspect that the natural tendencies to construct such rules… will cause experts to refine their rules rather than discard them. (Camerer and Johnson 1991, 210)

More generally, lack of awareness of the accuracy of expert judgment in linguistics is of particular concern, since linguists, often motivated to promote or criticize a particular theory, may be led to apply inferior strategies with more determination, a phenomenon Larrick has called the "lost pilot" effect ("I don't know where I'm going, but I'm making good time") (Larrick 2004, 321).

Finally, a word should be said about group decision making. Perhaps the collective judgment of multiple experts are more accurate than that of just one expert. If so, perhaps expert judgment in linguistics can be vindicated somewhat to the extent that we focus on those cases where multiple experts agree, and very few disagree. There are, however, at least four reasons for concern about collective judgments. First, such a strategy may require ignoring a large proportion of relevant cases, where controversy exists, and yet inferences are routinely drawn via expert judgment (with any collective judgment(s) remaining unknown). Second, to the extent that group expert judgment exists in linguistics, the aggregation of the individual judgments is typically highly informal, with one expert reporting her sense of what is "generally accepted" in the literature. There is much room for bias to creep into the aggregating linguist's judgment of how widespread a judgment is. Third, group judgment suffers from the same fundamental difficulty as individual expert judgment, namely that it is simply unknown how accurate the method is. Fourth, the evidence from the decision sciences community gives further reason for pessimism:

> perhaps the most insidious problem in groups is that people are unknowingly influenced by the public judgments of others. Especially under conditions of uncertainty, people are susceptible to anchoring on the judgments of others in forming their own judgments…. Shared training, shared experiences, and shared discussions all lead group members to hold a similar view of the world – *and* similar blind spots. (Larrick 2004, 326).

In summary, there are reasons for serious concern about the accuracy of the informal, verbal, holistic judgments of experts. This concern only increases when substantial feedback is unavailable, as it typically is in linguistics. Although they do not believe it, experts tend to misuse evidence, giving more weight to novel information, and growing increasingly miscalibrated with the facts as the issues get more difficult and as the experts' familiarity with quantitative methods decreases. Experts are typically unaware of these limitations, since they also exhibit much overconfidence with respect to both their beliefs about the plausibility of their favored theories and their beliefs about the accuracy of their judgmental powers. Inferences in linguistics typically take the form of expert judgments. Thus, it is reasonable to expect that as linguists combine and assess evidence like that in §2, their expert judgments will substantially distort the evidence's actual bearing on the linguistic theories in question.

4. The Conservative Backlash

In discussing the need for linguists to begin developing and using more explicit methods for analyzing and using their evidence, I have encountered a number of replies in defense of the methodological status quo. I will discuss four of these replies, and argue that they do not provide any support for this stance.

Reply 1: Disbelief. The first reply is an all too familiar reaction to the literature on expert judgment (e.g., Arkes 2003, Camerer and Johnson 1991). Whether it takes the form of simply ignoring the results, insisting that linguists are good at these kinds of judgments, or overtly denying the validity of the results, the first reply amounts to an outright denial that there is any problem with the current methods of drawing inferences. However, as we saw in §3, there is enormous evidence that expert linguistic judgments are, as is typical, highly inaccurate. It may be that linguists are in fact good at making expert judgments, but without evidence that linguists are radically unlike other experts in this regard, little credence should be given to such a hope.

Reply 2: Linguistics is more complicated. The second reply is also familiar. It usually takes the form of pointing out how much more complex the relevant linguistic evidence is than that used in the toy examples used to motivate the issue. It is uncontroversial that linguistic evidence is much more complicated and harder to understand than, say, two sequences of Heads and Tails. But it is utterly unclear why this fact should *support* the credibility of expert linguistic judgment; indeed, one would assume that it further *undermines* any such credibility. After all, when situations like these become more complex, human performance typically gets worse; rarely does it improve. As Dawes, Faust, and Meehl note (1989, 1672), if the rule for calculating the total at the supermarket gets more complex than simply adding prices, we wouldn't expect informal human judgment to get better:

> Suppose instead that the supermarket pricing rule were, 'Whenever both beef and fresh vegetables are involved, multiply the logarithm of 0.78 of the meat price by the square root of twice the vegetable price'; would the clerk and customer eyeball that any better? Worse, almost certainly.

In short, before this reply can have any force, we need reason to think that the kind of increased complexity found in linguistic data sets is likely to significantly *improve* expert judgment. It would be fascinating indeed if linguists turned out to be radically unlike other experts in this regard, but until reason is given for believing this surprising claim, it should not be endorsed.

Reply 3: Pessimism about improvements. The third reply is to cast doubt on whether it is possible (and feasible) to find any precise, explicit methods to replace the expert judgments. If there are no alternatives in the offing, then the claim that alternative methods should be used is an empty exhortation. Two things can be said about this reply. First, the quality of a method and the existence of alternatives to it are quite distinct issues. We would have little confidence in a report used Tarot cards to estimate the total mass of silicon on Pluto. Nor would our confidence in, or acceptance of, this method increase if the author explained that there was no known better way to construct the estimate. Barring reasons for thinking that Tarot cards could produce quality judgments on this topic, we would be faced with the choice of living without evidence regarding the silicon on Pluto, or seeking better methods. Although expert judgment enjoys a better reputation in linguistics than Tarot cards, §3 showed that there is good reason to think that the former method may also be highly inaccurate. In extreme cases it may not outperform a method based on Tarot cards. (In many of the studies cited above, experts performed at chance, or were even slightly negatively correlated with the normatively correct answers.) By making these unpalatable but plausible outcomes salient, my hope is to increase the sense of urgency in linguistics to develop and explore more explicit, analyzable methods for theorizing about human language.

Second, it is not true that there are no explicit methods that can reduce the reliance on expert linguistic judgment. Elsewhere (Johnson ms.) I have shown how some familiar statistical tools can be easily modified and combined to give a measure of the degree to which one linguistic theory has better empirical coverage than another. The issues are somewhat delicate, since competing linguistic theories often don't overlap perfectly in terms of what phenomena they each account for. Moreover, linguistic data arrives continuously, not as a single batch, and some of these data are associated with other members of the data set, and so they cannot be treated as "independent" points of empirical data. However, a considerable theoretical and mathematical grip can be gained using only the phi-coefficient and the logic of sequential probability ratio testing. The methods just mentioned do not, of course, address every issue where expert judgments are employed. Rather, they chip away at certain specific aspects of the general problem. But this "chipping away" procedure is the norm in the sciences – particularly those whose methodology is still undergoing substantial improvement. E.g., journals like *Biometrika*, *Econometrica*, and *The Journal of Mathematical Psychology* are filled with reports of new methods for attacking specific (often very narrow!) problems in their respective fields. These journals – along with faculty positions, conferences and societies dedicated to methodological issues – have been active for decades. It is, I submit, high time linguistics began following suit.

Reply 4: Analogies with other scientific methods. The final reply I consider involves appealing to methods used in other scientific disciplines. Linguists commonly compare their goals and methods to those of vision science, biology, physics, etc (cf. e.g., the papers in Chomsky, 2000, and Uriagereka 1998 esp. chap 1). In general, many of these comparisons are quite apt. Indeed, Chomsky's work in this area has been extremely important to the development of the theoretical foundations of linguistics. E.g., there is much to admire in Chomsky's discussions of why linguists need not worry about positing unobserved structure in their theories of language, reducing their theories to those of more "basic" disciplines, or employing only elements that are accessible to consciousness (Chomsky 2000, chaps. 4, 5; cf. also Johnson 2007). However, such discussions should not be overinterpreted. For instance, although some aspects of linguistic methodology are shared with physics (or astronomy, biology, chemistry, etc.), it does not follow that linguistics has "exhausted the methods of science" (Chomsky 1986, 252). It similarly does not follow that we are free to ignore the resulting "mere difference in degree only, not in kind" between the way the various fields implement these shared methodological aspects. Instead, this "difference in degree" can make all the difference in evaluating linguistic methodology.

To see this last point, notice that a proper assessment of the relationships between linguistic and other methodologies crucially involves assessing their differences as well as their similarities. A major difference between linguistics and the various (other) empirical sciences is that it is standard in the latter disciplines to use explicit quantitative methods for analyzing and assessing the data and its impact on various theories. Indeed, this is why calculus, algebra, and statistics are standard background training in these fields. There are many reasons why such methods are used in – and some would say are definitive of – the sciences. Not least among these reasons is the vast increase in accuracy and precision over simple human judgment. Further benefits include relative ease of analysis and insensitivity to contextual effects, biases, etc. By being explicit, the mathematical methods are themselves comparatively easy to study. By being subjected to such analyses, these method's strengths and weaknesses can be clarified, and the methods themselves can be improved. Moreover, by being explicit, the methods are insensitive to irrelevant features of the data or experimental situation in a way that humans are not. (For example, Dawes (1979) famously showed that a linear statistical model whose weights were *randomly* assigned to the variables could outperform a human expert who had access to the same information. This occurred because the model was consistent in its predictions, and did not vary from case to case, as the experts did.)

In addition to the comparison with the sciences, it is also instructive to compare linguistics to history, English literature, and other fields that are typically not considered "scientific". Here, too, there appear to be some rather strong similarities. These fields all gather evidence regarding various topics, and assemble it, primarily via the method of informal, holistic, verbal expert judgment, into a case that supports or undermines a theory or theories. What would happen if a linguist attempted to argue – with as much vigor as she argues that linguistics is a science – that linguistics is in fact one of the humanities? It is plausible that a very strong case could be made. What would this exercise show? The real payoff, I believe, would be to show that there is less benefit in defending linguistics as a "science" or a "humanities", and much more benefit in making a frank and open assessment of how the methods of contemporary linguistics work, where they are strong and where they are weak, and how they can be improved.

CONCLUSION

There is, to my knowledge, little reason for strong confidence regarding how linguistic evidence is routinely compiled and assessed. This paper raised some doubts about the accuracy of informal, verbal, holistic expert judgments in linguistics, and stressed the importance of accuracy in assembling diverse sources of evidence. Instead of using informal judgments, linguists should seek explicit ones that can be studied, analyzed and improved. Instead of using verbal judgments, linguists should seek precise, formal ones that are incapable of ambiguity, and whose biases (if any) must be worn on their sleeves, most likely in an exact quantitative format. Instead of using holistic, all-things-considered judgments, linguists should work to adopt the spirit of decision analysis, which is to divide and conquer: "Decompose a problem into simpler problems, get one's thinking straight in these simpler problems, paste these analyses together with a logical glue, and come out with a program for action for the complex problem" (Raiffa 1968, 271).

Hopefully, some of the examples and literature reviewed have brought out the importance to linguistics of the above recommendations. Unfortunately, however the evidence also suggests that overconfidence in expert judgments is remarkably resilient to the scientific evidence about them. It is probable that in linguistics as elsewhere, the first step – namely understanding and accepting the limited accuracy of expert judgments – will be the hardest to take.

References

Arkes, H. (2003). The nonuse of psychological research at two federal agencies. *Psychological science 14*, 1 – 6.

Armstrong, J. S (1980). Unintelligible management research and academic prestige. *Interfaces 10,* 80 – 86.

Bishop, M., and Trout, J. D. (2005). *Epistemology and the psychology of human judgment.* Oxford: OUP.

Camerer, C. F. and Johnson, E. J. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly?. In K. A. Ericsson and J. Smith (Eds.) *Toward a general theory of expertise: Prospects and limits* (pp. 195 – 217). Cambridge, MA: CUP.

Dawes, R. (1971). A case study of graduate admissions. *American psychologist 26*, 180 – 188.

Dawes, R. (1979). The robust beauty of improper linear models in decision making. *American psychologist 34*, 571 – 582.

Dawes, R. and Corrigan, B. (1974). Linear models in decision making. *Psychological bulletin 81*, 95 – 106.

Einhorn, Hillel J. (2000). Expert judgment: some necessary conditions and an example. In T. Connolly, H. Arkes, and K. Hammond (Eds.), *Judgment and decision making* (pp. 324 – 335). Cambridge, UK: CUP.

Faust, D., and Ziskin, J. (1988). The expert witness in psychology and psychiatry. *Science 241*, 31 – 35.

Gaeth, G., and Shanteau, J. (2000). Reducing the influence of irrelevant information on experienced decsion makers". In T. Connolly, H. Arkes, and K. Hammond (Eds.), *Judgment and decision making* (pp. 305 – 323). Cambridge, UK: CUP.

Griffin, D., and Brenner, L. (2004). Perspectives on probability judgment calibration. In D. J. Koehler, and N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 177 – 199). Oxford, UK: Blackwell.

Harvey, N. (1997). Confidence in judgment. *Trends in cognitive sciences, 1*, 78 – 82.

Henrion, M. and Fischhoff, B. (1986). Assessing uncertainty in physical constants. *American journal of physics, 54*, 791 – 798.

Johnson, K. (2007). The legacy of methodological dualism. *Mind and language 22*, 366 – 401.

D.J. Koehler, D. J., Brenner, L., and Griffin, D. (2000). The calibration of expert judgment. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases : The psychology of intuitive judgement* (pp. 686 – 715). Cambridge: CUP.

Kruger, J, and Dunning, D. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 1121 – 1134.

Landau, I. (2003). Movement out of control. *Linguistic inquiry, 34*, 471 – 498.

Larrick, R. P. (2004). Debiasing. In D. J. Koehler, and N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 16 – 337). Oxford, UK: Blackwell.

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306 – 334). Cambridge: CUP.

Lovie, A. D., and Lovie, P. (1986). The flat maximum effect and linear scoring models for prediction. *Journal of forecasting 5*, 159 – 168.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and review of the evidence*. Minneapolis: University of Minnesota Press.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of personality assessment, 50*, 370 – 375.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology, 2*, 175 – 220.

Nisbett, R. E. (1993). *Rules for reasoning*. Hillsdale, NJ: Erlbaum.

Phillips, J. K., Klein, G., and Sieck, W. R. (2004). Expertise in judgment and decision making: A case for training intuitive decision skills. In D. J. Koehler, and N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 297 – 315). Oxford, UK: Blackwell.

Raiffa, H. (1968). *Decision analysis*. Reading, MA: Addison-Wesley.

Shafir, E., and LeBoeuf, R. (2004). Context and conflict in multiattribute choice. In D. J. Koehler, and N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 341 – 359). Oxford, UK: Blackwell.

Tversky, A., and Kahneman, D. (1971). Belief in the law of small numbers. *psychological bulletin, 76*, 105 – 110.

Uriagereka, J. (1998). *Rhyme and reason: An introduction to minimalist syntax*. Cambridge, MA: MIT Press.